IEEE ComSoc IEEE Communications Society
IEEE Open Journal of the
Communications Society

Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/OJCOMS.2022.1234567

Real-time Immersive Aerial Video Streaming: A Comprehensive Survey, Benchmarking, and Open Challenges

Mohit K. Sharma[*], Ibrahim Farhat[*], Chen-Feng Liu[†], MEMBER, IEEE, Nassim Sehad[‡], Wassim Hamidouche[*], AND Mérouane Debbah[§], FELLOW, IEEE

[1]Technology Innovation Institute, Abu Dhabi, UAE
[2]Department of Informatics, New Jersey Institute of Technology, Newark, NJ, USA
[3]Aalto University, School of Electrical Engineering, Department of Information and Communications Engineering, Espoo, Finland
[4]Khalifa University, Abu Dhabi, UAE

CORRESPONDING AUTHOR: Mohit K. Sharma (e-mail: mohit.sharma@tii.ae).

ABSTRACT

Over the past decade, the use of Unmanned Aerial Vehicles (UAVs) has grown significantly due to their agility, maneuverability, and rapid deployability. An important application is the use of UAV-mounted 360-degree cameras for *real-time streaming of Omnidirectional Videos (ODVs)*, enabling immersive experiences with up to six Degrees-of-freedom (6DoF) for applications like remote surveillance and gaming. However, streaming high-resolution ODVs with low latency (below 1 second) over an air-to-ground (A2G) wireless channel faces challenges due to its inherent *non-stationarity*, impacting the Quality-of-experience (QoE). Limited onboard energy availability and energy consumption variability based on flight parameters add to the complexity. This paper conducts a thorough survey of challenges and research efforts in UAV-based immersive video streaming. First, we outline the end-to-end 360-degree video transmission pipeline, covering coding, packaging, and streaming with a focus on standardization for device and service interoperability. Next, we review the research on optimizing video streaming over UAV-to-ground wireless channels, and present a real testbed demonstrating 360-degree video streaming from a UAV with remote control over a 5G network. To assess performance, a high-resolution 360-degree video dataset captured from UAVs under different conditions is introduced. Encoding schemes like AVC/H.264, HEVC/H.265, VVC/H.266, VP9, and AV1 are evaluated for encoding latency and QoE. Results show that HEVC's hardware implementation achieves a good QoE-latency trade-off, while AV1's software implementation provides superior QoE. The paper concludes with discussions on open challenges and future directions for efficient and low-latency immersive video streaming via UAVs.

INDEX TERMS 360° video, extended reality, low latency, real-time streaming, low latency, UAV.

I. INTRODUCTION

IMMERSIVE video technology enables users to experience a quasi-realistic virtual environment, fostering engagement and a sense of presence in a digital space. Various visual media modalities, such as volumetric, light field, and Omnidirectional Video (ODV), have emerged as viable options for delivering an immersive viewing experience [1]. Among these, ODV, commonly known as 360-degree video, has gained widespread popularity due to the availability of acquisition and display devices, and standardization efforts ensuring interoperability. To enhance immersion, interaction with the user is crucial. This interaction can involve head movements (roll, yaw, and pitch), mouse/keyboard controls, or in the case of viewing on a smartphone, the viewing angle can be controlled by moving the device in space, providing a visual experience of up to three Degrees-of-freedom (3DoF). However, one of the main limitations of ODV is the absence of motion parallax, which refers to the relative position of objects changing based on the viewer's position relative to the object. This can lead to discomfort and motion sickness for users.

To address this limitation, a potential solution is to employ a 360° camera mounted on a Unmanned Aerial Vehicle (UAV). This combination offers enhanced flexibility and mobility, allowing users to explore the environment and move around objects within the scene. By leveraging the mobility provided by the UAV, in addition to 360° video, a viewing experience of up to six Degrees-of-freedom (6DoF) can be achieved. This advancement holds promise for diverse applications like remote video surveillance, scientific exploration, autonomous manufacturing assistance, agricultural monitoring, and more. However, to fully realize the potential of these applications, it is crucial to maintain a seamless and responsive interaction between the user and the UAV by ensuring a more natural viewing experience with accurate control. This requires ODV to be delivered with high Quality-of-experience (QoE), to ensure a truly immersive experience through real-time control of the UAV. Specifically, the high-quality 4K resolution videos need to be transmitted with ultra-low End-to-end (E2E) latency (preferably below 1 sec. [2]). However, achieving these metrics over contemporary 5G networks is highly challenging due to the higher data volume of ODVs, compared to conventional Two-dimensional (2D) videos. For instance, an High-efficiency Video Voding (HEVC)-encoded 8K (ultra-high-definition) video typically requires target bitrates ranging from 20-80 Mbps [3], significantly exceeding the typical throughput of 20 Mbps for UAVs when operating in the presence of ground users [4], [5]. Furthermore, achieving Glass-to-glass (G2G) latency of under one second is inherently challenging. This is because a 30 frames-per-second video encoded with a Group-of-pictures (GOP) size larger than 16 inherently incurs a G2G latency of at least one second. However, reducing the size of the GOP negatively impacts compression efficiency. Further, the intrinsic *non-stationarity* of the UAV-to-ground wireless channel and limited computational and energy resources of UAVs further amplify these challenges for UAV-based real-time immersive video streaming.

Addressing the above challenges will require efforts to enhance not only the communication for UAVs and develop adaptive and low-complexity schemes for 360° video encoding and streaming, but also to consider the *interplay between these two design issues*. It is important to note that the design approach of a system for real-time streaming from a UAV mounted 360° camera needs to be completely different compared to a ground-based immersive video streaming system. This is because of the inherent dependence of the air-to-ground (A2G) wireless channel on the UAV trajectory and its location in the space, leading to *non-stationarity* and a fundamentally different behavior compared to terrestrial channels between a base station and a ground-based user. On the other hand, the existing design of UAV-based 2D streaming [6] cannot be directly adapted, due to the interactive nature of immersive streaming and higher data rate requirements. In addition, the interplay between the UAV trajectory, onboard energy availability, computation power, encoding, and communications need to be analyzed carefully to deliver a truly immersive experience.

In addition, we note that the existing 360° video datasets captured from a ground-based camera do not capture essential characteristics of UAV-based 360° videos, e.g., vertical motion. Because the encoding parameters critically depend on the video content, the performance of standard video encoders needs to be re-evaluated afresh on aerial 360° videos to understand their QoE and latency performance.

The rest of this paper is organized as follows. In the following section, we present a review of the existing literature on this topic, and in Sec. III we describe the main components of the ODV streaming chain, including acquisition, encoding, packaging, rendering, and optimization. Then, the key performance metrics and wireless optimization techniques for UAV-based 360° video streaming are presented in Sections IV and V, respectively. In Sec. VI, we present a review of Third Generation Partnership Project (3GPP) activities relevant to real-time streaming of ODVs from a UAV platform. Further, in Sec. VII, the proposed aerial 360° video dataset is presented, and then benchmarking results and analysis of software and hardware encoders of five video standards are provided in Sec. VIII. Next, the challenges of ODV streaming from a UAV platform are discussed in Sec. IX. Finally, Sec. X concludes the paper.

II. RELATED WORK & CONTRIBUTIONS

In Table 1, we present a summary of recent efforts [6]–[22] surveying state-of-the-art research on communication for UAVs and immersive streaming. The literature in Table 1 can be broadly classified into two categories: covering the communication aspects of UAVs and the streaming of 360° videos. The authors in [7]–[9] presented a comprehensive survey of challenges and fundamental tradeoffs in designing wireless networks involving the UAVs. In particular, Mozaffari *et al.* [7] described analytical frameworks and tools to address design challenges, and Hayat *et al.* [9] surveyed the quality of service, connectivity, safety, and other general networking requirements for unmanned aircraft systems in civilian applications. Baltaci *et al.* [10] reviewed the connectivity requirements for aerial vehicles, especially for piloting applications, and advocated achieving these stringent connectivity requirements through multi-technology heterogeneous networks. In [8], [11], the authors evaluated various enabling 6G technologies, highlighting the benefits, drawbacks, and challenges in their integration into 6G wireless network with UAVs. The authors in [12] surveyed the channel models for air-to-ground and air-to-air UAV communication.

To address high data rate requirements for UAVs, Xiao *et al.* [13] reviewed antenna structures and channel models for millimeter wave (mmWave). Furthermore, the technologies and solutions for UAV-connected mmWave cellular networks and mmWave-UAV ad hoc networks were discussed. The authors in [14] and [15] reviewed the methods for communication and trajectory co-design. In addition, Zeng *et al.* [14] surveyed techniques to deal with the issues on air-to-ground interference in cellular communication with UAVs. Fotouhi *et al.* [16] also surveyed the interference issues in serving aerial users with the existing terrestrial

TABLE 1. Overview of the State of the Art

Category	Summary		
	References	Survey Focus	Aerial ODV Streaming
Communication for UAVs	[7]	Three-dimensional (3D)-deployment, Performance and energy efficiency analysis, Channel modeling	No
	[8], [11]	The potential of advanced technologies for UAV's integration into 6G networks: Intelligent reflecting surfaces, Short-packet communication, Integrated communication and sensing	No
	[9]	Connectivity requirements for aerial communications	No
	[10]	Characteristics & Requirements of UAV networks: • Quality-of-service (QoS) & data rate requirements • Network-relevant mission parameters • Connectivity, safety, privacy, security, and scalability	No
	[12]	UAV communications channel model, link budget analysis	No
	[13]	mmWave technologies for UAV communications	No
	[14]	Trade-offs between QoS, size, weight, power constraints, and 3D mobility	No
	[15]	3D obstacle avoidance mechanisms	No
	[16]	Interference issues, standardization activities, and cyber-security	No
	[17], [18]	Standardization, Aerial experimentation and research platform	No
	[19]	Joint design of communications, computation, and control for performance improvement	No
360° video streaming	[6]	Video streaming (2D) from aerial platforms	No
	[20]	Challenges in on-demand and live 360° streaming, standardization activities, and architectures	No
	[21]	Compression and coding for 360° streaming, Network issues in Virtual Reality (VR) streaming	No
	[22]	Data model for 360° video, challenges and approaches for creating and distribution of 360° videos	No

Base Stations (BSs), along with potential solutions proposed by standardization bodies. In addition, they reviewed the ongoing prototyping, testbed activities, and regulatory efforts to manage the commercial use of UAVs, along with cyber-physical security of UAV-assisted cellular communication. In [17], Marojevic *et al.* presented an architecture and research platform for aerial experimentation with advanced wireless communications, which facilitates experimental research in controlled yet production-like environments. In [18], Abdalla *et al.* surveyed the ongoing 3GPP standardization activities for enabling networked UAVs, requirements, envisaged architecture, and services provided by UAVs. The authors in [19] studied the UAV networks from the perspective of cyber-physical systems and considered the joint design of communication, computation, and control to improve the performance of UAV networks. We note that most of the existing research efforts do not explicitly cover the aforementioned unique issues, described in the previous section, relevant to immersive video streaming from a UAV platform.

On the other hand, the work in [6], [20], [21] surveyed the adaptive streaming techniques for 360° videos. Yaqoob *et al.* [20] reviewed the adaptive 360° video streaming approaches that dynamically adjust the size and quality of the viewport. In addition, they surveyed the standardization efforts for 360° video streaming, highlighting the main research challenges such as viewport prediction, QoE assessment, and low latency streaming for both the on-demand and live 360° video streaming. Further, [21] surveyed the Field-of-view (FoV) prediction methods, along with compression, and coding schemes for reducing the bandwidth required for streaming immersive videos. In addition, they reviewed caching strategies and datasets for immersive video streaming. The work in [6] focused on 2D video streaming from an aerial platform. In particular, they surveyed the works using Artificial Intelligence (AI)-based techniques to enhance the video streaming performance. While these works provide key insights into various aspects of immersive video streaming from a ground-based platform, they fail to capture the unique characteristics and trade-offs of the aerial immersive video streaming systems.

In this work, we present a thorough survey of key trade-offs, challenges, and research efforts in UAV-based immersive video streaming. In addition, we benchmark the existing video encoding schemes for their encoding latency and QoE, using a high-resolution 360-degree video dataset captured from UAVs under different conditions. Our contributions are the following:

- We present a comprehensive review of existing video streaming efforts from a UAV, and provide key insights into the design trade-offs.
- We present a new 360° video dataset, captured from a UAV in diverse acquisition conditions.
- Assess the coding efficiency and complexity of software and hardware encoders of five video standards and formats for immersive 360° video streaming.
- We highlight the open challenges related to ODV streaming from a UAV.

This is the first paper surveying the key trade-offs, research efforts, and open design challenges for UAV-based real-time immersive streaming. In addition, the presented dataset of 360° videos captured from UAV is the first in the field and will aid research efforts in joint optimization of communication and encoding schemes for real-time immersive streaming of aerial 360° videos. In the following section, we describe the main blocks of an ODV streaming pipeline.

III. OMNIDIRECTIONAL VIDEO STREAMING

An omnidirectional visual signal is presented in a spherical space with angular coordinates: the azimuth angle $\phi \in [\pi, -\pi]$, and the elevation or polar angle $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, assuming a unit sphere (radius $r = 1$) for acquisition and rendering. The sphere's origin represents the viewing reference that captures the light coming from all directions. Since the human visual system has a limited field of view, at a time a user cannot view the entire 360° content in its spherical representation. Instead, only a portion of the sphere, known as the "viewport", is displayed, which is an image tangent to the sphere. Initial streaming approaches, termed *viewport-independent streaming*, involved transmitting the entire 360-degree video content at high quality, allowing users to extract the desired viewport based on their head position, with low latency. However, it is a bandwidth-intensive solution, requiring over 100 Mbps to transmit an 8K resolution video at high quality [23]. This is inefficient since the end user only observes a small portion (approximately 15%) of the ODV. To address this, more advanced techniques have been proposed to transmit only a portion of the sphere, corresponding to the current viewport. Due to their superior bandwidth efficiency, *viewport-dependent* strategies have gained wide adoption at the projection (projection-based) and encoding (tile-based) stages. As shown in Figure 1, ODV streaming strategies can be broadly categorized as either viewport-dependent or viewport-independent, depending on whether the FoV is considered in the optimization process or not.

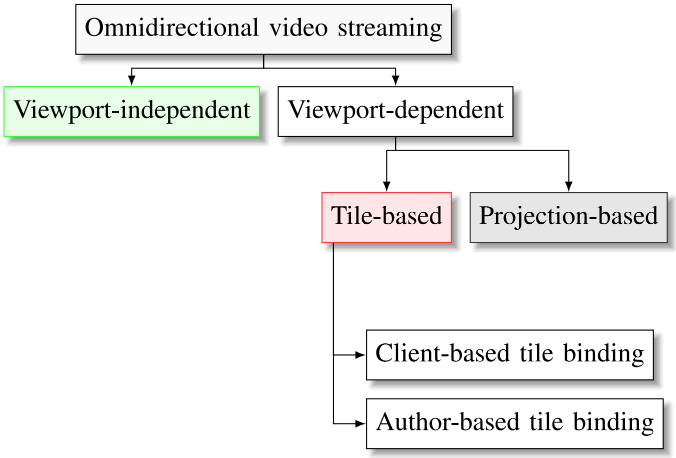

Omnidirectional video streaming
Viewport-independent
Viewport-dependent
Tile-based
Projection-based
Client-based tile binding
Author-based tile binding

FIGURE 1. ODV streaming strategies.

In the following, we provide an overview of the architecture of the E2E ODV streaming pipeline, illustrated in Figure 2. We briefly describe the technology used at each stage to deliver ODV to the end user, highlighting the features included to support viewport-dependent streaming.

A. ACQUISITION AND PREPROCESSING

In practice, an omnidirectional visual signal is captured using a multi-view wide-angle acquisition system, often utilizing fish-eye lenses. Since a single eye-fish camera can only capture a partial sphere, combining multiple acquisitions from such cameras allows for complete sphere coverage through the process of *stitching* the images [24]. However, the *stitching* operation introduces two main challenges. The first challenge involves blending and wrapping non-overlapping captured images, while also addressing inconsistencies in illumination and color that may arise after stitching. The second challenge arises when dealing with video signals, as the camera sensors need to be perfectly synchronized.

The omnidirectional visual signal in spherical representation is mapped over another surface during the preprocessing stage to facilitate further processing after acquisition. At a high level, the mapping schemes differ in terms of the geometry of the surface to be mapped. The most commonly used mapping technique is Equirectangular Projection (ERP), which is particularly well-suited for production and contribution purposes, and uniformly maps the pixels on

TABLE 2. List of Acronyms

Acronym Definition	Acronym Definition
2D : Two Dimensional	**LTE**: Long Term Evolution
3D : Three Dimensional	**M2P**: Motion-to-photon
3GPP: Third Generation Partnership Project	**MCTS**: Motion-constrained Tile Set
6DoF: Six Degrees-of-freedom	**MIMO**: Multi-input Multi-output
A2G: Air-to-ground	**MLLM**: Multimodal Large Language Model
AI: Artificial Intelligence	**MPEG**: Motion-picture Expert Group
ANN: Artificial Neural Network	**MV**: Motion Vector
AP: Access Point	**NLoS**: Non Line-of-sight
AVC: Advanced Video Coding	**NS**: Network Slicing
AV1: AOMedia Video 1	**ODV**: Omnidirectional Video
BD-rate: Bjøntegaard-Delta rate	**OMAF**: Omnidirectional Media Format
BS: Base Station	**PSNR**: Peak Signal-to-noise Ratio
CMP: Cube Map Projection	**QoE**: Quality-of-experience
CNN: Convolution Neural Network	**QoS**: Quality-of-service
DASH: Dynamic Adaptive Streaming over HTTP	**RSRP**: Reference Signal Received Power
E2E: End-to-end	**RTP**: Real-time Transport Protocol
EM: Eye Movement	**RWP**: Region-wise Packing
ERP: Equirectangular Projection	**S-PSNR**: Spherical PSNR
FoV: Field-of-View	**SRTP**: Secure Real-time Transport Protocol
FPV: First Person View	**SSIM**: Structural Similarity Index Measure
G2A: Glass-to-algorithm	**SVC**: Scalable Video Coding
G2G: Glass-to-glass	**TSP**: Truncated Square Pyramid
GOP: Group-of-pictures	**UAV**: Unmanned Aerial Vehicle
GPU: Graphic Processing Unit	**UHD**: Ultra-high Definition
HEVC: High-efficiency Video Coding	**VMAF**: Video Multi-method Assessment Fusion
HM: Head Movement	**VR**: Virtual Reality
HMD: Head-mounted Display	**VS**: Viewport Specific
ISOBMFF: ISO Base Media File Format	**VVC**: Versatile Video Coding
LoS: Line-of-sight	**WebRTC**: Web Real-time Communication
LLM: Large Language Model	**WLAN**: Wireless Local Area Network

the sphere over a rectangular plane. More advanced mapping techniques, such as Cube Map Projection (CMP) and Truncated Square Pyramid (TSP), map the spherical signal over the six faces of a cube and square-based pyramid with four triangular faces [25], respectively. Notably, compared to ERP, CMP and TSP offer enhanced coding efficiency, achieving bitrate savings of 25% and 80%, respectively, making them more suitable for distribution purposes [26].

On the other hand, *dynamic projection methods*, such as pyramidal projection and its refined version, offset cubic projection [27] facilitate viewport-dependent streaming by modulating the pixel density depending on the viewing direction. Offset cubic projection allocates higher pixel density and better quality near the offset direction which corresponds to the user's viewing direction. Another solution proposed in [28] is oriented projection for real-time 360-degree video streaming, which allocates more pixels in the projected frame to areas on the sphere that are close to a target pixel-concentration orientation.

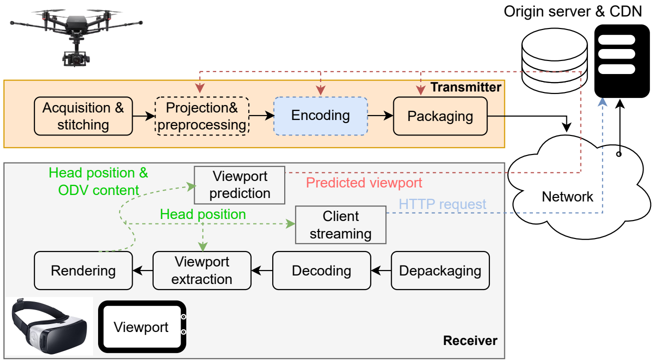


FIGURE 2. ODV E2E streaming pipeline. Note that, the HTTP request by WebRTC client is used for signaling.

B. ENCODING

After mapping the sphere in a 2D plane[1], ODV content is encoded in practice by conventional 2D

[1] For ease of exposition, we describe the ODV pipeline with the ERP.

video standards such as Advanced Video Coding (AVC)/H.264 [29], HEVC/H.265 [30], Versatile Video Coding (VVC)/H.266 [31], as well as VP9 and AOMedia video 1 (AV1) video formats. In particular, tailored coding tools are integrated into the HEVC/H.265 and VVC/H.266 standards to enhance the ODV coding efficiency and enable advanced streaming features, improving the user's QoE.

1) HEVC/H.265 Tools for ODV

The HEVC/H.265 leverages the tile concept, where the mapped pixels are subdivided into small non-overlapping rectangular regions, to facilitate the viewport-dependent streaming. The tile concept enables independent and parallel encoding/decoding of rectangular regions within the picture. By breaking the dependency of context prediction in arithmetic encoding and intra-prediction, tiles allow for efficient processing and coding of specific regions [32]. Additionally, the tile boundaries also enable the possibility of disabling in-loop filters, further enhancing the flexibility of the encoding process. Moreover, the introduction of the Motion-constrained Tile Set (MCTS) technique in HEVC/H.265, along with supplemental enhancement information messages, extends the tile concept to the sequence of frames. This technique restricts the Motion Vectors (MVs) to a selected set of tiles in the reference picture, thereby enabling the fetching and decoding of only the tiles within the displayed viewport during ODV streaming. This approach significantly improves the user's QoE by delivering high-quality content while efficiently utilizing bandwidth. However, the limitation of restricting MVs within a set of tiles in the reference picture can decrease coding efficiency. To overcome this, the literature proposes non-normative solutions that enhance inter-prediction by utilizing the base layer as a reference in the scalable HEVC extension [33]. Alternatively, Bidgoli *et al.* [34] propose an enhanced intra-prediction technique with fine granularity random access capability, allowing end-users to request specific parts of the stream while ensuring efficient intra-coding. Furthermore, in the context of spherical bitrate allocation, a new entropy equilibrium optimization strategy is proposed in [35]. This strategy derives the Lagrangian multiplier at the block level, which is used in rate-distortion optimization. The proposed solution, evaluated with ERP and CMP, demonstrates significant bitrate gains when compared to the HEVC reference software encoder [35].

2) VVC/H.266 Tools for ODV

The VVC/H.266 standard introduces several advancements for efficient encoding of ODV content, including the ability to signal the used projection technique and the definition of tailored coding tools [31]. In the case of 360-degree representation and ERP mapping, objects can span across the left and right picture boundaries continuously. Consequently, in VVC/H.266, inter-prediction samples may wrap around from the opposite left or right boundary when MVs point outside the coded area. Additionally, virtual boundaries are defined to skip in-loop filters across edges. For CMP, where cube maps may exhibit content discontinuities, virtual boundaries can be signaled to disable in-loop filtering and prevent artifacts arising from non-homogeneous boundaries. Furthermore, VVC/H.266 introduces the concept of subpictures, which allows for the extraction of independent rectangular regions within the picture, specifically designed for viewport-dependent VVC streaming applications. Subpictures offer two critical improvements over the previous MCTS concept. Firstly, subpictures enable MVs to refer to blocks outside the subpicture, and padding at subpicture boundaries is permitted, similar to picture boundaries. This facilitates higher coding efficiency compared to the tight motion constraints applied in MCTS. Secondly, a need to rewrite slice headers when extracting a sequence of subpictures to build a new VVC/H.266 compliant bitstream is eliminated, streamlining the encoding process [31].

In addition to standard encoders, some non-normative techniques are also proposed in the literature, e.g., encoding the ODV content in spherical representation to prevent projection distortions, leading to higher coding efficiency.

3) Learning-based coding for ODV

Machine learning techniques have been extensively investigated in the literature to optimize and improve the coding efficiency of ODV content. In [36], a Convolution Neural Network (CNN) was trained to learn the rotation of the sphere, resulting in an improvement in the coding efficiency. This rotation is applied as a pre-processing step along the spherical axis before projection, leading to different rotations of the cube map. Experimental results demonstrate that incorporating rotation prediction achieve a significant coding gain of 8% to 10% with a prediction accuracy of 80%.

Similar to conventional video standards, learning-based video codecs can encode ODV content after its projection onto a 2D plane. Initially, the 2D representation is transformed into a compact latent space using an analysis transform based on an Artificial Neural Network (ANN). The resulting latent representation is then encoded with a lossless entropy encoder to construct the bitstream. At the decoder side, a synthesis transform, also based on an ANN, reconstructs a version of the input 2D representation from the received bitstream. Moreover, the hyperparameters of the latent space entropy distribution, such as mean and variance, are encoded using an auto-encoder and utilized by the encoder and decoder to enhance the performance of the entropy encoder [37].

C. STREAMING PROTOCOLS

Various packaging protocols can be employed for streaming ODV content, depending on the specific application and end-user requirements concerning video quality, latency, and advanced functionalities provided by the protocol [38]. In the following, we outline the key features of two widely utilized streaming protocols: Omnidirectional Media Format (OMAF) and Web Real-time Communication (WebRTC). For further details, readers are encouraged to refer to overview papers on OMAF [39] and WebRTC [40].

1) OMAF

The ISO/IEC 23090-2 standard, also known as OMAF, is a system standard developed by the Motion Picture Experts Group (MPEG) to ensure device and service interoperability for storing and streaming omnidirectional media content. This includes various forms of media such as 360° images and videos, spatial audio, and associated text. The initial version of the standard, completed in October 2017, provides fundamental tools for streaming 360° images and videos, enabling a 3DoF viewing experience. In the subsequent release of the standard in October 2020, the second version introduced additional tools to support more advanced features. These features include enhanced viewport-dependent streaming, overlay capabilities, and the ability to stream multiple viewpoints, marking the initial steps towards achieving a 6DoF viewing experience, desired for UAV-based real-time immersive streaming.

The specifications of OMAF are organized into three main modules: content authoring, delivery, and player. Furthermore, these specifications serve as extensions to the ISO Base Media File Format (ISOBMFF) and Dynamic Adaptive Streaming over HTTP (DASH), ensuring backward compatibility with conventional 2D media formats. OMAF supports three types of omnidirectional visual signal representations: projected, mesh, and fish-eye. Each of these formats requires specific pre-processing for encoding and post-processing for rendering and display. Among the projected formats, OMAF includes support for two widely used projection algorithms: ERP and CMP. Additionally, OMAF incorporates a Region-wise Packing (RWP) operation, which allows for optional pre-processing operations before encoding. These operations include resizing, repositioning, rotation by 90°, 180°, and 270°, as well as vertical and horizontal mirroring of specific rectangular regions. RWP serves various purposes, such as signaling the exact coverage of a partial spherical representation, generating Viewport-specific (VS) video, enhancing coding efficiency, or compensating for over-sampling in the pole areas of ERP. The RWP metadata indicates the applied operations to the player, which then performs inverse operations to map the regions of the decoded picture back into the projected picture. This ensures proper rendering and display of the content, aligning with the intended transformations specified by the RWP.

The OMAF standard supports both viewport-dependent\independent streaming profiles, as outlined in [27]. The viewport-dependent ODV streaming profile of OMAF enables the selection of segments covering the user's viewport at high quality and other segments at lower quality and bitrate. This approach allows for more efficient utilization of network bandwidth, resulting in an improved user experience. Viewport-dependent ODV streaming can be achieved through two methods: VS and tile-based streaming. In the Viewport-specific approach, multiple VSs are created and signaled, each encoding different viewports at high quality. Users can select the appropriate VS stream based on their viewing orientation. The OMAF region-wise quality ranking metadata can be used to signal the quality of different regions in the sphere. On the other hand, in the tile-based configuration, the ODV is divided into independent rectangular regions called *tiles*. Following the projection stage, the ODV is encoded into tiles representing different quality representations. The end user can then request the tiles covering the viewport at high quality, while the remaining area tiles can be requested at a lower quality. Each tile only depends on the co-located tile in the sequence and can be decoded independently of other tiles. There are two alternatives for encoding video in independent regions. The first method utilizes the HEVC tile concept, where tiles are grouped into motion-constrained slices known as Motion-constrained Tile Sets. This profile employs HEVC encoding to achieve low-quality coverage of the entire 360-degree video, while high-quality sub-pictures are encoded to cover specific regions of the video. The second method, applicable to AVC which does not support tiles, partitions the video into sub-picture sequences, each representing a spatial subset of the original sequence. These sub-picture sequences are then encoded with motion constraints and merged into tiles in a single bit-stream. Each tile or sub-picture sequence is stored in its respective track. Additionally, tiles can be encoded in different bitrates and resolutions, allowing users to select the optimal combination of tiles based on viewing orientation, available bandwidth, and decoding capability.

In total, the OMAF standard specifies six video media profiles that define the type of video representation and the supported video standard with its associated levels. For example, the "HEVC-based viewport-independent" profile uses the ERP representation and is constrained to HEVC Main 10 profile level 5.1. This level limits the spatial resolution to 4K (4096 × 2160). However, the "unconstrained HEVC-based viewport-independent" profile, introduced in the second edition, supports all HEVC Main 10 profile levels, thus increasing the decoding capacity and display resolution. Furthermore, there are already several open-source implementations available that support the first[2] edition of the OMAF standard. Further, some tools of the OMAF second edition have been demonstrated in [41], [42].

[2] NOKIA: https://github.com/nokiatech/omaf, Fraunhofer HHI: https://github.com/fraunhoferhhi/omaf.js, Intel Open Visual Cloud: https://github.com/OpenVisualCloud/Immersive-Video-Sample.

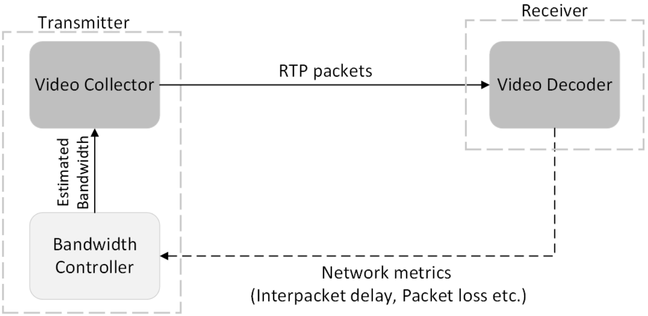


FIGURE 3. **WebRTC block diagram.**

2) WebRTC

The WebRTC framework is an open-source solution specifically designed to facilitate real-time and low-latency video transmission. As shown in Figure 3, within the WebRTC transmitter, the “video collector” module performs video encoding and encapsulates the encoded video frames into Real-time Transport Protocol (RTP) packets. These packets are subsequently transmitted using the *secure real-time transport protocol*. On the receiver side, relevant information regarding the received RTP packets is collected, and this information is relayed back to the “video collector” through the transport-wide feedback message of the *real-time transport control protocol*. The “bandwidth controller” module, located within the “video collector,” utilizes these control messages to compute essential network metrics such as inter-packet delay variation, queuing delay, and packet loss. These metrics play a crucial role in determining the target bitrate, which is then employed by the rate control module of the video encoder. The rate control module dynamically adjusts the encoding parameters, such as the quantization parameter and resolution, based on the target bitrate requirements. Although, unlike OMAF, the standard WebRTC implementation does not offer explicit tools for transmitting immersive video, it has gained significant popularity for real-time and ultra-low latency ODV transmission by treating 360° video representation as a conventional 2D video [28], [43]. Additionally, viewport-dependent streaming can be effectively supported by incorporating a combination of high-resolution and low-resolution tiles. This approach optimizes bandwidth utilization while ensuring high quality within the FoV and maintaining a low Motion-to-photon (M2P) latency [44].

D. RENDERING AND DISPLAY

The limited FoV of the human visual system prevents the end users from visualizing the entire 360° content in the spherical representation. Therefore, only a portion of the sphere (i.e., an image tangent to the sphere called *viewport*) is displayed. The viewport acts as a window through which the viewer can observe a segment of the entire spherical video. The positioning and size of the viewport are dynamically adjusted based on the viewer’s head and gaze orientation, which are tracked in real-time by the Head-mounted Display (HMD)’s sensors. By continuously tracking the user’s head movements and adjusting the viewport accordingly, the mobility provided by the UAV along with the 360° video facilitates a viewing experience that supports up to 6DoF. Nonetheless, to achieve seamless rendering and display of remotely captured UAV sequences, with accurate viewport adaptation, the end-to-end latency in the downlink control channel needs to be ultra-low to adjust the UAV position depending on the head and eye tracking data.

As described in the previous subsections, the spherical video content is transmitted in an equirectangular format, where the video frame represents a flattened sphere. A critical step before rendering is to effectively project this flat image back onto a sphere within the VR environment. This transformation requires meticulous geometrical adjustments to ensure that the spherical illusion is maintained without visible artifacts or distortion. In terms of display technology, modern VR headsets utilize dual organic light-emitting diode-based display or liquid crystal display panels that offer fast response times and high refresh rates, essential for maintaining immersion and reducing motion sickness. Each eye views its display, and the combined effect of this binocular display creates a stereoscopic effect, enhancing the content’s depth and fullness of presence. To optimize the viewer’s experience, modern devices employ rendering techniques that prioritize the resolution and update rate of the area within the viewport. This method, often referred to as *foveated rendering* [45], reduces the graphical fidelity in the peripheral vision outside the immediate area of focus, thus allowing for higher frame rates and improved resolution where it is most crucial—typically in direct line of sight. These features enable a more natural viewing experience and better remote control, effective for user interaction.

In the following section, we present the key performance metrics for a UAV-based immersive video streaming system and benchmark the technologies discussed in this section.

IV. UAV-BASED REAL-TIME IMMERSIVE VIDEO STREAMING: PERFORMANCE METRICS

This section describes the key performance metrics for real-time UAV-based immersive video streaming systems. The discussion encompasses three essential aspects: latency, video quality, and UAV energy consumption. We highlight key trade-offs involved in optimizing these metrics and their impact on the design choices. In addition, we benchmark the various technologies discussed in the previous section.

A. LATENCY

The latency in video transmission from a UAV significantly impacts the user’s QoE in 360-degree video streaming. It is captured using metrics such as E2E latency, M2P latency, and motion-to-high resolution latency, defined below.

End-to-end latency: In a point-to-point real-time video transmission, E2E latency plays a vital role in ensuring a seamless and immersive experience. It represents the total delay from event capture by the sensor to actuator response, including processing and transmission latency. The E2E latency between the camera and user's display is often referred to as G2G latency. It measures the difference between time instances when the photons of an event first pass through the camera lens and when the event is displayed to the viewer. Another metric, termed as Glass-to-algorithm (G2A) latency, represents the time gap between the photon corresponding to an event passing through the camera lens and the availability of the first image corresponding to that event for processing before display. G2A latency is crucial in applications utilizing computer vision algorithms for tasks such as control, object detection, segmentation, and viewport prediction. Figure 2 provides an overview of G2G latency and its relationship to G2A latency. At a high level, the total G2G latency encompasses the delay between the input at the acquisition and stitching block and the output of the rendering block. It comprises network latency as well as latency originating from video processing components at both the transmitter and client sides. The overall G2G latency can be expressed as the sum of delays incurred during camera acquisition, encoding, network transmission, decoding, and display processing. Notably, G2A latency can be derived from G2G latency by subtracting the latency introduced during the rendering and display processes.

Table 3 presents a breakdown of G2G latency for a state-of-the-art WebRTC-based implementation of an ODV streaming pipeline [46]. This pipeline transmits $8K$ resolution 360° videos captured using an Insta 360 camera to a Samsung S10 client. The latency breakdown in Table 3 highlights that the acquisition and stitching process, along with the encoder, contributes to approximately 80% of the total G2G latency. We note that the total G2G latency shown in Table 3 also includes the transmission latency, incurred over the network. It is important to note that *the latency introduced at the transmitter scales proportionally with the video resolution and the frame rate.*

Based on the preceding discussion, it can be deduced that reducing latency entails reducing the number of processed pixels across the ODV streaming pipeline, which is primarily determined by the frame rate and resolution. Additionally, higher frame rates and quality necessitate increased transmission rates, resulting in higher overheads in transmission delay and transmit power. On the other hand, when the encoding bitrate does not dynamically adapt to fluctuations in the wireless channel conditions, the queuing delay increases due to the generation of more data than the instantaneous wireless channel capacity, leading to an increased E2E delay. As a result, efforts to minimize latency have a direct impact on video quality. Therefore, the design of wireless communications for UAV-based ODV streaming predominantly revolves around maximizing video quality while adhering to a latency constraint, typically imposed as a *delay outage probability* constraint representing the probability of packet delay exceeding a predefined delay budget. Note that, the delay outage probability constraint only encompasses queuing and transmission delays, focusing on a portion of the overall G2G delay by ignoring the latency introduced during the preprocessing, encoding, and packaging of the ODV data.

TABLE 3. Glass-to-Glass Latency Brake up [46]

Block	Latency (ms)	
Transmitter	Live Streaming	503
	FFMpeg Decoder	568
	360° stitching	28.5
	HEVC encoder	406
	Video packetizer	1.9
	Total latency at transmitter	**1508**
Client	RTP Packet	79
	Decoder	34
	Renderer	14
	Total latency at client	**127**
Total G2G latency		**1745-1856**

Motion-to-photon latency: For *viewport-dependent streaming*, the user's quality of service is better captured by the latency metrics such as M2P latency and motion-to-high resolution latency. M2P latency measures the delay required to display the new viewport corresponding to the user's updated viewing direction after head movement. It measures the time needed to request and render the viewport aligned with the user's viewing direction. The specific streaming approach and the technology of the HMD can influence the motion-to-photon latency. Additionally, recent work presented in [47] demonstrates the potential of utilizing head motion prediction algorithms at the end user's side to significantly reduce the M2P latency. These algorithms can effectively anticipate the user's Head Movements (HMs) and optimize the rendering process accordingly.

B. ODV QUALITY

In addition to the latency metrics described above, an end user's Quality-of-experience (QoE) is primarily determined by the perceived video quality. For 2D videos, widely used full-reference objective quality metrics include Peak Signal-to-noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Video Multi-method Assessment Fusion (VMAF). These metrics provide a comprehensive assessment of the perceived quality by comparing the original and reconstructed videos. However, for 360-degree video content, specialized quality metrics have been proposed to account for the unique geometrical distortions introduced by the spherical representation. Notable examples include Spherical PSNR (S-PSNR) and weighted to spherically uniform PSNR, which are *full-reference* objective quality metrics specifically developed for 360-degree video content [48]. However, to assess video quality in the presence of imminent frame drops due to adverse wireless channel conditions, novel no-reference objective quality metrics must be developed. This is crucial as the majority of existing AI-based no-reference quality metrics rely on data availability, which is not always guaranteed in scenarios with frame drops.

As noted earlier, video quality is influenced by various factors, including encoding bitrate, frame resolution, frame rate, and the characteristics of the air-to-ground wireless channel. Generally, higher quality and lower distortion can be achieved by using a higher bitrate (or resolution) while exploiting the favorable channel conditions. However, it is important to note that bitrate selection not only affects video quality but also impacts latency, via increased processing and queuing delay. In addition, a higher bitrate necessitates a more stringent throughput requirement, posing challenges for efficient wireless resource allocation. Therefore, the system design is characterized by a trade-off between reconstruction quality/distortion and bitrate selection, which is impacted by the requirement for optimal provisioning of wireless resources to meet the selected video bitrate.

Note that, in contemporary systems, the enhancement in video quality not only increases the latency, but also the energy consumption in the pre-processing and encoding stages [49]. In particular, the energy consumption increases in direct proportion to the increase in the number of pixels, frames, and bitrate used for encoding. Considering the limited energy available[3] on a UAV, the QoE implicitly affects the UAV flight time. In the following, we elaborate on this trade-off.

C. UAV ENERGY CONSUMPTION/ FLIGHT TIME

The energy consumed by a UAV during movement is referred to as “propulsion energy,” which is influenced by the UAV’s velocity and acceleration. Additionally, when the UAV hovers at a fixed position while streaming the video, it consumes “hovering energy” [52], [53]. On the other hand, as discussed below, the A2G channel between the UAV and the ground user directly depends on the UAV’s position in the 3D space which, in turn, also determines the energy consumption. For instance, the small-scale fading component of the UAV-ground wireless channel can be modeled as an “angle-dependent Rician fading channel” with the Rician factor directly proportional to the UAV-ground elevation angle [54]. This model captures the fact that as the elevation angle increases, the UAV-ground link tends to experience less scattering, resulting in a larger Line-of-sight (LoS) component. In addition, the large-scale fading component, which includes path-loss and shadowing, depends not only on the 3D locations of the UAV and the ground user but also on the geographic distribution of buildings. In urban areas, the signal propagation of a UAV flying at a lower altitude may be obstructed by buildings, leading to the shadowing effect [55]. In contrast, when the UAV transmits at a higher altitude, it only experiences path loss without any shadowing. However, conducting a comprehensive path-loss measurement for a wide geographic area is infeasible. Therefore, a generic probabilistic A2G channel model that statistically incorporates both LoS and Non Line-of-sight (NLoS) large-scale fading is used [56]. In this model, the probability of experiencing LoS path increases as the UAV raises its altitude or moves closer to the ground user horizontally.

Note that the energy consumed in data transmission forms an important component of the onboard energy consumption of a UAV, which also includes the transmit power. Further, the transmit power affects both the latency as well as the quality of the received video, as it determines the probability of the successful transmission of data packets. While the power consumption for communications is notably lower than that for hovering and propulsion, it is not insignificant [57] and thus warrants optimization. Overall, due to the limited on-board energy, the total power/energy consumption – including the power consumption of the onboard Graphics Processing Unit (GPU) used for encoding and pre-processing – becomes a crucial factor in the design of UAV-based real-time 360° video streaming systems, significantly impacting the design choices.

Thus, the trajectory and position of a UAV affect not only its energy consumption but also the quality of the transmitted video. Hence, in the deployment and trajectory design of UAVs for video streaming, the distinctive features of the air-to-ground channel, as well as the propulsion and hovering energy consumption needs to be accounted.

D. BENCHMARKING AND OPTIMIZATION

The metrics to be optimized for UAV-based real-time streaming consist of perceived video quality, flight time, required bandwidth, and various latency measures (e.g., E2E, M2P, or motion-to-high-resolution/quality latency). Low M2P latency is particularly important to minimize user discomfort when changing the displayed viewport while achieving low E2E latency is crucial to enable accurate remote control, especially during high-speed flying.

As discussed in Figure 1, ODV streaming strategies can be categorized as either viewport-dependent or viewport-independent, depending on whether the FoV is considered in the optimization process or not. Table 4 benchmarks the performance of 360° video streaming strategies, de-

[3] Majority of the UAVs are equipped with a single battery to power the drone, LIDAR, and the CPU [50], [51] and the battery size is limited by the weight consideration.

TABLE 4. **Performance of Streaming Approaches.**

	Bandwidth	Latency	Encoding complexity
Viewport-indep.	●○○	●●●	●●●
Projection-based	●●○	●●○	●●○
Tile-based	●●●	●●○	●○○

Performance metrics: High ≡ ●●●, Average ≡ ●●○, Low ≡ ●○○

scribed in Sec. III, with respect to required bandwidth, M2P latency, and encoding complexity[4]. As can be observed from Table 4, although the tile-based encoding incurs low-bandwidth usages and moderate M2P latency, encoding ODV tiles in multiple representations, each with different rate-quality characteristics, leads to high encoder computational complexity and E2E latency. To improve this, the VR Industry Forum guidelines [60] introduced the HEVC-based FoV Enhanced Video Profile. This profile employs HEVC encoding to achieve low-quality coverage of the entire 360-degree video, while high-quality sub-pictures are encoded to cover specific regions of the video. Each bitstream is then encapsulated within a track compliant with the HEVC-based viewport-dependent OMAF profile. The player can subsequently request the bitstream covering the viewport in high quality, along with the low-quality bitstream representing the entire 360-degree coverage. Moreover, in live scenarios, the low-quality stream can be transmitted via multicast, allowing for more efficient bandwidth utilization, while maintaining ultra-low motion-to-photon latency.

Furthermore, the prediction of end-user HMs can be leveraged to enhance QoE by assigning higher fetching priority to tiles within the predicted viewport. This "human-centric" streaming approach focuses on optimizing the user experience, in contrast to the "system-centric" approach that prioritizes overall system performance without considering user behavior. The design can be categorized as single-user or cross-user, with the latter considering the behavior of multiple users in predicting the viewport. These techniques rely on accurate *viewport prediction models*, which are used to optimize the streaming system. In [61], the potential of predicting HMs for optimizing 360-degree video streaming over cellular networks was demonstrated, resulting in up to 80% network bandwidth savings. This approach has been adopted by several research papers and commercial products, aiming to optimize network and computational resources and provide a highly immersive experience [62].

The following section provides a detailed exploration of these design challenges and state-of-the-art, by providing a comprehensive survey of the research efforts in the wireless community to address the challenges in live immersive streaming from a UAV.

V. QoE OPTIMIZATION & PERFORMANCE EVALUATION

In this section, we present a comprehensive survey of the research on optimization and performance evaluation of UAV-based real-time video streaming systems. First, we review the work focusing on QoE maximization through optimal wireless resource allocation, and next, we describe the research that also leverages trajectory optimization as an additional 'degree-of-freedom' for optimizing the performance. Further, we present an overview of work on evaluating the performance of these systems in diverse settings.

A. OPTIMIZATION

1) Wireless Resource Allocation for QoE Maximization

The inherent randomness of wireless channels poses a significant challenge in achieving a high QoE, as varying channel conditions result in unpredictable latency, which leads to interrupted or choppy video streaming. Maximizing QoE is generally approached as a problem of maximizing PSNR by optimizing wireless resource allocation, including transmit power, rate, or bandwidth, while adhering to the wireless network and UAV-imposed constraints. In this section, we survey state-of-the-art advancements in this area.

In [63], Xia *et al.* utilized the internal sensor data of the UAV for adaptive bitrate selection. They leveraged location, velocity, and acceleration information to predict future throughput and proactively select the video bitrate accordingly. The performance was evaluated using a DJI Matrice 100 drone with an attached Android smartphone in an outdoor environment, communicating with a laptop on the ground using the IEEE 802.11n protocol. The simulations demonstrated that the selected bitrates effectively adapted to future throughput, maintaining relatively stable video bitrates over time, resulting in a seamless video viewing experience despite channel fluctuations. In another study, Muzaffar *et al.* [64] studied a multicast video streaming framework where a UAV delivers video to ground users. The proposed approach incorporated feedback from the users to dynamically adjust the transmission rate and video bitrate. The performance evaluation was conducted using the AscTec Pelican drone equipped with a Logitech C920 camera and employing the IEEE 802.11a protocol and AVC/H.264 video format, investigated throughput, packet loss, and delay. The rate-adaptation approach demonstrated improvements in throughput, latency, and packet loss compared to a constant transmission rate and bitrate baseline, resulting in up to 30% PSNR gain. These works represent significant advancements in enhancing QoE through adaptive bitrate selection and rate control mechanisms, showcasing the potential of optimizing wireless communications for UAV-based video streaming.

[4] Quantifying the encoding complexity of a video involves measuring various factors that contribute to the computational complexity and resources required to encode the video. Several factors impact the encoding complexity: including spatial complexity, temporal complexity, bitrate and resolution, and quantization parameter. Further, the encoding complexity can be quantified in terms of metrics such as encoding time and computational load. We refer the reader to [58], [59] for a further discussion on this.

In [65], the authors considered a multi-UAV setup where UAVs competed for transmission rates by incurring a cost to obtain higher rates. Each UAV aimed to maximize its utility, comprising PSNR and cost, by selecting a transmission rate within the network capacity budget. The authors designed a rate allocation algorithm using game theory to address the rate competition among UAVs. Compared to the equal rate allocation baseline, the proposed algorithm increased network utility while considering video quality requirements.

Another line of work, e.g., [66], [67], attempts to maximize the PSNR by using a Scalable Video Coding (SVC) based video transmission. In SVC, the video is encoded into a base layer and N enhancement layers. If the nth enhancement layer (or quality) is selected for the streamed video, the base layer and all lower enhancement layers, i.e., $1, \cdots, n-1$, have to be delivered along with the nth layer [66]. Note that, more enhancement layers lead to a better quality of the received video, i.e., the higher PSNR, but require more transmit power at the UAV. In [66], Zhang *et al*. considered a system where a UAV transmits video to a terrestrial BS with SVC. The objective was to maximize the energy efficiency subject to the *delay outage probability* constraint, i.e., the probability that packet delay exceeds a predetermined delay budget. Energy efficiency is defined as the ratio of the PSNR to the total power. The optimal solution jointly determined the number of enhancement layers and transmit power. In contrast with the baseline, which randomly selects the number of layers and power, the proposed approach improved the energy efficiency by 40% and decreased the delay outage probability from 0.3 to 0.05. The work [67] studied a system in which the base and enhancement layers of the SVC video are sent from a terrestrial BS and the UAV BSs, with storage and computation capabilities, to the ground users. Each layer of the video can be served by either the terrestrial BS or a UAV BS, i.e., a user can obtain the layers of the video from multiple BSs. The computation capabilities at the BSs can be used for video processing, e.g., encoding the video's base layer and enhancement layers. In addition, the UAVs without the storage and computation capabilities act as relays to help the transmission from the terrestrial BS to the users. Since the number of enhancement layers affects the video quality, the users desire more enhancement layers. By optimizing the transmit power and allocated bandwidth of the BS and UAVs, the number of enhancement layers for the users, the video layer assignment (i.e., from which BS), and the 2D deployment of the UAVs, the objective in [67] was to maximize the sum of all users' QoE metric, e.g., normalized PSNR, subject to the constraint on the transmission and computation delays. The proposed approach achieved 15% better QoE, i.e., received video quality improvement, than a baseline, where the video layers for the user originate from a single BS delivering the highest throughput. In contrast with the other baselines in which the video layers for all users originate from the terrestrial BS, and the video transmission is helped by the UAV relays, the proposed approach achieved 68% QoE enhancement. However, it is important to note that due to its *high computational complexity* and lack of support by broad-based consumer devices, the SVC based approach is not preferable for real-time video transmission.

We note that all the above-discussed work focused only on the transmission of 2D videos from UAVs. In contrast to this, Hu et *al.* [68] conducted a numerical analysis of a UAV-based ODV streaming system, where ground users request specific video tiles within their FoV from the UAV. The UAVs then transmit the requested tiles to the users via associated Access Points (APs) which act as decode-and-forward relays. These APs collaboratively broadcast the video data to the corresponding users. The objective of their approach was to maximize the PSNR by scheduling time slots to the UAVs and associating them with the APs. The proposed approach yielded an enhancement in PSNR compared to baselines where APs worked either totally independently or collaboratively.

2) UAV Deployment and Trajectory Design

Along with wireless resource allocation, such as transmit power and bandwidth, the maneuverability of UAVs offers an additional dimension for enhancingstreaming performance, by improving both throughput and latency. By optimizing the UAV's location or trajectory in 3D space, both energy consumption and wireless channel conditions can be improved.

Guo *et al.* [69] focused on the 3D trajectory design of a UAV deployed to inspect multiple facilities and transmit real-time video to a control center. The objective was to minimize the total energy consumption associated with propulsion and hovering. The trajectory between successive facilities directly impacted propulsion energy, while hovering energy depended on the inspection time at each facility, determined by video bitrate and transmission latency. Therefore, a trajectory planning algorithm was proposed in [69] to minimize total energy consumption, assuming a fixed video bitrate. Simulation results demonstrated that the proposed algorithm significantly reduced the UAV's energy consumption and flight time. The work in [70], undertook joint optimization of trajectory and resources, e.g., time slots, transmit power, and transmission rate, for a UAV-based video delivery to multiple ground users. The trajectory design took into account the propulsion energy consumption. The authors formulated the user's utility as the normalized transmission rate relative to a predetermined bitrate (considering fairness among users). They aimed to maximize the lowest time-averaged utility among all users by jointly designing trajectories and allocating wireless resources. The proposed approach outperformed three baselines: trajectory optimization, wireless resource optimization, and no optimization. It achieved up to a 3-fold increase in transmission rate. Building upon the work in [69], Bur *et al.* [71] considered a scenario of collaborative inspection of a fire area by multiple UAV

TABLE 5. Summary of the QoE Optimization Research

	Performance Metrics	**Optimization Variables**
Xia *et al.* [63]	QoE (re-buffering time, jitter, and quality)	Transmission rate, video bitrate
Muzaffar *et al.* [64] & He *et al.* [65]	Video quality	Transmission rate
Zhang *et al.* [66] & Liang *et al.* [67]	Video quality normalized to energy consumed	Transmit power, number of enhancement layers, and bandwidth allocated
Hu *et al.* [68]	Video quality	Scheduling and association with AP
Guo *et al.* [69]	Energy consumption	3-D trajectory design
Zhan *et al.* [70]	Energy consumption and flight time	Trajectory design, transmit power, and time slots
Burhanuddin *et al.* [71]	Data transmission rate and latency	Transmit power, trajectory, and bitrate
Chakareski *et al.* [72]	Video quality	Transmit power, trajectory, and coding
Khan *et al.* [73]	Video quality	Trajectory, UAV deployment, and bitrate allocation

users, with the inspection videos sent to a UAV-BS. The optimization involved the transmit power of all UAVs, 3D trajectories of UAV users, and dynamic bitrates of the inspection videos transmitted by the users. The focus was on QoE maximization, which accounted for transmission delay violation and the normalized transmission rate based on the selected video bitrate. Additionally, the transmission rate was constrained to be greater than the selected video bitrate, considering the trajectories and transmit power of the UAVs. The proposed approach supported the transmission of 720p and 1080p videos with an average delay of 0.05 ms, whereas a greedy approach relying on immediate QoE decisions only supported 140p videos with an average delay of 1.2 ms. Overall, these studies highlight the importance of jointly optimizing UAV trajectories and resource allocation to enhance video streaming performance, achieving energy efficiency, reduced delay, and improved QoE. In [72], the authors developed a dynamic placement strategy for multiple UAVs to maximize the expected immersion fidelity for a scene of interest. The objective was to minimize the overall reconstruction error of all users by optimizing transmit power and source-channel coding.

Furthermore, Khan *et al.* [73] investigated a UAV-to-UAV communication network where UAVs collaboratively streamed video to a ground server. Their approach involved utilizing dual paths for transmitting SVC video with one enhancement layer. The base layer is sent directly from a UAV to the ground server via a radio frequency link, while the enhancement layer is relayed to the server by neighboring UAVs using free-space-optical links. The objective was to minimize distortion in the received video by jointly optimizing the bitrates of the base and enhancement layers, the routing path, and UAVs deployment. The optimization was subject to a constraint on propulsion energy consumption and the channel capacity's bitrate limitations. The proposed approach achieved an average PSNR gain of 6 dB, compared to a baseline approach that used dual paths with only radio frequency links, without optimizing the routing path and UAV deployment. In summary, Khan *et al.* explored UAV-to-UAV communication networks, demonstrating the benefits of jointly optimizing routing paths, UAV deployment, and bitrate allocation for enhanced video streaming performance.

TABLE 6. Summary of Testbed & Measurement Activities

Work	**Measurement Objectives**
Stornig *et al.* [74]	Impact of UAV mobility on video quality & latency
Zhou *et al.* [75]	To evaluate transmission delay, packet loss probability of control command and video data
Jin *et al.* [76]	Study G2G delay and transmisison rates over 4G and 5G networks
Taleb *et al.* [77]	Measurement of PSNR and G2G delay over 4G and 5G networks
Qazi *et al.* [78] & Sinha *et al.* [79]	Throughput, delay, and packet loss evaluation in various indoor and outdoor network configurations
Naveed *et al.* [80]	To study the relationship between RSRP and throughput, and its impact on video quality
Liu *et al.* [81] & Nihei *et al.* [82]	Evaluate the effect of multipath streaming on E2E delay
Yu *et al.* [83]	Throughput and energy performance for 4K uncompressed video transmission over mm-wave networks

B. TESTBED & MEASUREMENT ACTIVITIES

In the following, we survey the testbed setups and measurement activities focused on evaluating the video quality and the network performance, characterized by throughput and latency, for video transmission from a UAV.

Stornig et al. [74] employed the ns-3 network simulator to study E2E delays and video quality metrics (PSNR and

SSIM) for video streaming over 4G networks. They modeled the UAV's 3D trajectory using a Gauss-Markov mobility model, and the video traffic was simulated using the MPEG-4 formats with the Evalvid application. The impact of UAV mobility on latency performance was thoroughly examined. Simulation results indicated that approximately two-thirds of frames were received with good or excellent quality, while 27% of frames in regular mobility and 30% of frames in high mobility exhibited inferior quality. Moreover, the average PSNR and SSIM values for the received video were 33 dB and 0.945, respectively, indicating good quality.

In the testbed presented in [75], a DJI Matrice 100 drone equipped with the Quectel EC25 Long Term Evolution (LTE) module and a Raspberry Pi camera were utilized. A computer with a USRP B210 radio frequency unit served as the BS, connected to the UAV remote controller via a wireline connection. The experiments were conducted indoors using the AVC/H.264 encoded videos. Various metrics were evaluated, including transmission delay, packet loss probability of control commands, and video data throughput. The results demonstrated that when the control command was updated less than 40 times per second, the command delivery experienced a 20 ms transmission delay without any packet loss. Also, the average delay and throughput for 480p and 720p video resolutions ranged from 1.5 s to 5.5 s, and from 2 Mbps to 9 Mbps, respectively. In [76], the authors evaluated the performance of a testbed equipped with the Huawei MH5000 5G module, operating in an outdoor environment. The transmission rates for streaming 1080p video in HEVC/H.265 format over 4G and 5G networks were measured at 16 Mbps and 97 Mbps, respectively. The G2G delays were evaluated as 3 s and 1.2 s for the respective networks. Additionally, the E2E delay of control command delivery was measured to be 30 ms in the 5G network.

In UAV teleportation, an operator at a remote location guides the UAV to accomplish tasks, over a live video feed. This requires simultaneous uplink streaming of real-time video and downlink delivery of control commands. Targeting these applications, the work in [77] implemented an *immersive* UAV control testbed using the Oculus Quest 2 HMD to control UAV movement and FoV over 4G, 5G, and WiFi networks. The Insta360 One X camera captured 360° video, and streaming rates of 2 Mbps to 8 Mbps were considered. Various delay metrics were evaluated: G2G delay, glass-to-reaction-to-execution delay, and sensor reaction delay. The G2G delay ranged from 0.595 sec. to 0.985 sec., the glass-to-reaction-to-execution delay ranged from 0.89 sec. to 1.38 sec., and the sensor reaction delay ranged from 0.67 sec. to 1.12 sec., as the streaming rate varied from 2 Mbps to 8 Mbps. The control command transmission delay was measured at 138 ms, 103 ms, and 88 ms for 4G, 5G, and WiFi networks, respectively. Additionally, the PSNR of the received video for 720p and 4K resolutions ranged from 30 to 47 dB.

In [78], [80], the network simulator ns-3 and Evalvid application were used to investigate the performance of MP4 format video transmission from the UAV to the BS in 4G networks. The study in [78] primarily focused on throughput investigation in both outdoor and indoor environments. In the outdoor scenario, the average throughput achieved by a static macrocell UAV was found to be 60 kbps, which decreased to 20 kbps as the UAV moved at speeds ranging from 1 to 5 m/s. In the indoor environment, the improvement in throughput was more significant for multi-story buildings with an increased number of deployed femtocell BSs. In a related work, Sinha *et al.* [79] leveraged network simulator ns-2.29 to evaluate the throughput, packet loss, packet retransmission, and E2E delay performance of video streaming between UAVs and from a UAV to the ground control station, in different network configurations, including wireless local area network (WLAN), WLAN router, WiFi hotspot, and WiFi Direct. Results indicated that WiFi Direct achieved the best performance for all metrics, followed by the WiFi hotspot, while the WLAN network exhibited the poorest performance in all considered metrics.

Naveed *et al.* [80] explored the relationship between the Reference Signal Received Power (RSRP) and throughput. Their findings revealed that as the RSRP varied from -110 dBm to -75 dBm, the UAV achieved video streaming throughput ranging between 2 Kbps and 80 Kbps. Additionally, the authors evaluated the received video quality using PSNR and SSIM scores under various wireless channel conditions. The PSNR scores were observed to be 49.41 dB, 35.42 dB, and 24.31 dB in the best, good, and poor channel conditions, respectively. Similarly, the SSIM scores were found to be 0.99, 0.63, and 0.35 in the respective channel conditions. Furthermore, the effects of various channel conditions on video quality were visually highlighted.

The performance evaluation of multi-path video streaming in 4G networks was conducted by Liu & Jiang [81], and Nihei *et al.* [82]. In the testbed presented in [81], video data was transmitted from dual devices inside the UAV to a smartphone. The dual-stream approach employed in this study demonstrated the capability to reduce the E2E delay to approximately 50 ms. In an independent study, Nihei *et al.* [82] tested the multi-path video streaming method in 4G networks, for forest fire surveillance, by distributing the video data over two 4G mobile network operators in Indonesia. The objective of data splitting was to minimize the average E2E delay. The experimental setup involved the use of a DJI Spreading Wings S800 drone equipped with a Raspberry Pi. Outdoor experiments were conducted using the AVC/H.264 format encoded videos. Visual illustrations provided in the study showcased the quality improvement achieved with the multi-path method. The performance of 60 GHz mmWave for video transmission was evaluated by Yu *et al.* [83]. In their experiment, conducted in an outdoor environment, a 4K uncompressed video was transmitted from the UAV to a nearby server to offload further computations.

The testbed achieved a throughput of 1.65 Gbps, and the results indicated that offloading computations to the server enabled the UAV to save 271.8 watts in computations at the expense of 4.1 watts for mmWave communication.

Based on the aforementioned results, it can be concluded that the design of wireless systems for UAV-based video streaming can vary depending on the specific wireless network architectures employed. Each network architecture comes with its restrictions, advantages, overheads, and hardware requirements, leading to diverse performance outcomes. These evaluation outcomes can also serve as guidance for selecting an appropriate network architecture, depending on the application requirements of UAV-based video streaming. It is worth noting that while most of the studies discussed in this section focused on non-real-time video streaming, they offer valuable insights into the design of UAV-based real-time ODV streaming. For example, the work by Yu *et al.* [83] emphasizes the importance of joint communications, computation, and control design for UAV-based real-time video streaming. Similarly, the results presented in [81], [82] demonstrate the effectiveness of multi-path streaming in significantly reducing E2E delays.

VI. OVERVIEW OF 3GPP STANDARDIZATION ACTIVITIES

In this section, we survey the relevant standardization activities conducted by 3GPP. The standardization activities related to UAV-based immersive video streaming within the 3GPP can be divided into two main categories. The first category focuses on the integration of UAVs with cellular networks, while the second category includes efforts on 5G support for media streaming applications, such as augmented reality, VR, and real-time communication. In the rest of the section, we provide an overview of the recent advancements and state-of-the-art in these two areas.

A. COMMUNICATION for UAVs

To evaluate the potential of LTE networks in supporting UAVs through cellular connectivity, the 3GPP initiated the Release 15 study in March 2017 [84]. The findings of this study are documented in TR 36.777 [85]. The study revealed that the LoS signal propagation in UAV communications increases the likelihood of severe interference in both uplink and downlink scenarios. Consequently, various interference detection and mitigation solutions were proposed as study and work items. Additionally, solutions related to mobility information management and aerial user identification were put forth. In Release 16, the focus shifted towards investigating the feasibility of remotely identifying UAVs [86]. In Release 17, 3GPP further addressed the operational 5G support for UAVs by providing functionalities for UAV authentication, authorization, and tracking [87]. Moreover, it allows for command and control authorization.

TABLE 7. Summary of 3GPP Release 18 Activities for Supporting Media Streaming over 5G Networks

3GPP Document	Focus
TS 26.501 [89]	5G Media Streaming (5GMS); General description and architecture
TS 26.506 [90]	5G real-time media communication architecture
TS 26.522 [91]	5G real-time media transport protocol configuration
TS 26.803 [92]	Study on 5G media streaming extensions for edge processing
TR 26.927 [93]	Artificial intelligence and machine learning in 5G media services

B. SUPPORT FOR MEDIA STREAMING OVER 5G

The support for VR over wireless networks was investigated in 3GPP Release 15, and conclusions are documented in TR 26.918 [88]. This report aimed to identify the potential gaps and use cases for facilitating VR services over wireless networks. Further, Release 17 TS 26.118 introduced operation points, such as resolution and color mappings, and defined media profiles for the distribution of VR content. To address the challenges associated with real-time immersive media streaming, Release 18 of 3GPP is currently investigating several relevant issues. For a comprehensive overview of the activities under Release 18, refer to Table 7.

Based on the above discussion, we note that the development of UAV-based real-time immersive streaming system is still in its infancy. For instance, the choice of the most suitable video encoder is still not clear from the available set of standard encoders. A key reason for this is the lack of a standard evaluation approach to provide a common benchmark for the developed algorithms. Towards this, in the following section, we present a dataset consisting of 360° videos captured from a UAV under various flying conditions.

VII. AERIAL ODV DATASET

As discussed earlier, the utilization of visual attention and saliency information can provide valuable insights into human visual scene analysis patterns. Visual attention and saliency information can be derived by analyzing viewers' HM and Eye Movement (EM) during video playback. This knowledge can be harnessed to develop effective encoding and streaming methods. However, it is important to note that for real-time video transmission, the HM and EM data can only be collected causally. Therefore, it needs to be collected in real-time and leveraged in an *online* manner to enhance the performance of real-time ODV streaming from UAV. On the other hand, there is no existing dataset containing aerial ODVs captured from a UAV. In this section, we present a survey of ODV datasets containing EM and HM information. In addition, we introduce a new dataset that we have curated for this study, containing ODVs captured from a UAV.

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2024.3455763

TABLE 8. **Summary of Existing Datasets**

Dataset	Resolution	Frame rate	Dimension	Description
EyeTrackUAV2 [94]	1280 × 720 and 720 × 420	30 fps	2D	Eye tracking data
AVS1K [95]	1280 × 720	30 fps	2D	Eye tracking data
WinesLab [96]	1080 × 1920	30 fps	360°	Videos recorded using both handheld and UAV mounted camera
360 Track [97]	3840 × 2160	30 fps	360°	Includes the ground truth for tracking
Proposed	3840 × 2160	30 − 50 fps	360°	Table 9

In the literature, several ODV datasets contain EM and HM information of viewers [98]. For better understanding the user behavior while watching ODVs, these datasets categorize the ODVs, based on the number of moving objects and camera motion, and include users' feedback about their viewing experience [99]. In contrast, [100] classified the videos based on their genre, such as documentaries, movies, etc. The majority of these datasets consist of videos with 3DoF which makes them less suitable for learning the user viewing pattern for a UAV-based ODV streaming. Indeed, inferences obtained using ODV with 3DoF may not be applicable for video transmission platforms with 6DoF, such as UAV-based ODV transmission. This raises the need to develop novel datasets of ODVs captured using UAVs. In the following, we briefly survey the existing datasets based on the videos captured from UAVs.

While many datasets in the literature include images and 2D videos captured by UAVs for applications such as remote sensing and navigation, only a limited number of publicly available datasets capture EM and HM information for UAV-recorded videos, with only one dataset currently accessible [96]. Similarly, there is only one dataset available for UAV-based 360° videos. We summarize these datasets in Table 8. The EyeTrackUAV2 dataset [94] collects binocular gaze information from 30 viewers watching 43 2D videos under both free viewing and task conditions. The AVS1K dataset comprises ground truth salient object regions for 1000 videos observed by 24 viewers in free viewing conditions. The WinesLab dataset contains *eleven* 360° videos, seven of which were recorded by a pedestrian using a handheld camera, and the remaining four were captured using a drone-mounted camera in various surroundings and lighting conditions. The 360Track dataset consists of *nine* 360° videos with manually marked ground truth positions of salient objects. In the following, we describe our dataset of *aerial* 360° videos, presented in Table 9.

The dataset presented in Table 9 comprises a total of *ten* 360-degree videos. The resolution of all videos, except for "FreeStyleParaGliding," is 3840 × 1920, while "FreeStyleParaGliding" has resolution 5120× 2560. Each video sequence in the dataset has a length of 40 seconds. All the videos, except "DubaiVertical" and "AbuDhabiCity," have a frame rate of 30 frames per second (fps), whereas "DubaiVertical" and "AbuDhabiCity" consist of 50 fps. The dataset consists of five outdoor videos, one sports video, and one video recorded in nighttime conditions. The "NorthPoleTrip" video captures motion in the azimuth plane, while the "DubaiVertical" video captures motion in the elevation.

In the following section, we use the above dataset to study the suitability of standard video encoders for real-time 360° streaming from a UAV. Also, we present our testbed for 360° streaming from a UAV.

VIII. BENCHMARK AND ANALYSIS

In this section, we first perform a comprehensive performance benchmarking of five video coding standards and formats (i.e., AVC/H.264, HEVC/H.265, VVC/H.266, VP9, and AV1) through their software implementations: libx264, libx265, Fraunhofer versatile video encoder (VVenC), libvpx-vp9, and libsvtav1, respectively. We also considered two NVIDIA hardware encoders, namely hevc_nvenc, and avc_nvenc, for the AVC/H.264 and HEVC/H.265, respectively. Next, we present a real-time drone ODV streaming testbed, employing a hardware AVC/H.264 encoder and WebRTC streaming protocol, for remote UAV control and navigation with a 6DoF viewing experience.

A. CODING AND COMPLEXITY PERFORMANCE

In this section, we evaluate the coding and latency performance of the above-mentioned software and hardware encoders on the video sequences contained in the dataset. Table 11 lists the used hardware and software encoder libraries for the five standards and formats. All the encoders are configured in their fastest preset, targeting live 360° video streaming applications. The encoding was conducted on a DELL precision 7820 tower workstation, equipped with an Intel Xeon CPU with 8 cores, running at a maximum frequency of 3.9 GHz, and a NVIDIA RTX A5000 GPU. The quality of decoded 360° videos is assessed using three objective quality metrics: Spherical PSNR (S-PSNR), SSIM, and VMAF. The videos are encoded at four practical UAV target bitrates of 1.5 Mbps, 3 Mbps, 4.5 Mbps, and 5.8 Mbps [101], enabling the computation of the Bjøntegaard-Delta rate (BD-rate) performance. The BD-rate gives the average bitrate saving or loss compared to the anchor encoder over the four considered bitrates.

Figures 4(a), 4(b), and 4(c) provide the average quality performance of the encoders on the proposed dataset, using

TABLE 9. Summary of Our Dataset

Sequence Name	Spatial Resolution	#Frames	Frame rate (fps)	Scene Feature
PetraJordan	3840 × 1920	1200	30	Outdoor
CapeTownCityPenorama	3840 × 1920	1200	30	Outdoor
CapeTownCityBeach	3840 × 1920	1200	30	Outdoor
CapeTownCityGarden	3840 × 1920	1200	30	Outdoor
CapeTownCitySquare	3840 × 1920	1200	30	Outdoor
FreeStyleParaGliding	5120 × 1920	1200	30	Sports
StPetersBergMuseum	3840 × 1920	1200	30	Night
NorthPoleTrip	3840 × 1920	1200	30	Motion
DubaiVertical	3840 × 1920	2000	50	Vertical Motion
AbuDhabiCity	3840 × 1920	2000	50	City Panorama

TABLE 10. Specifications of Workstation Used for Simulations

CPU	Intel Xeon Silver
#cores	8
Max Freq (GHz)	3.9
RAM	32 GB
SSD	256 GB
GPU	NVIDIA Ampere RTX A5000
Operating System	Ubuntu 20.04

TABLE 11. Video Encoder SW/HW Libraries

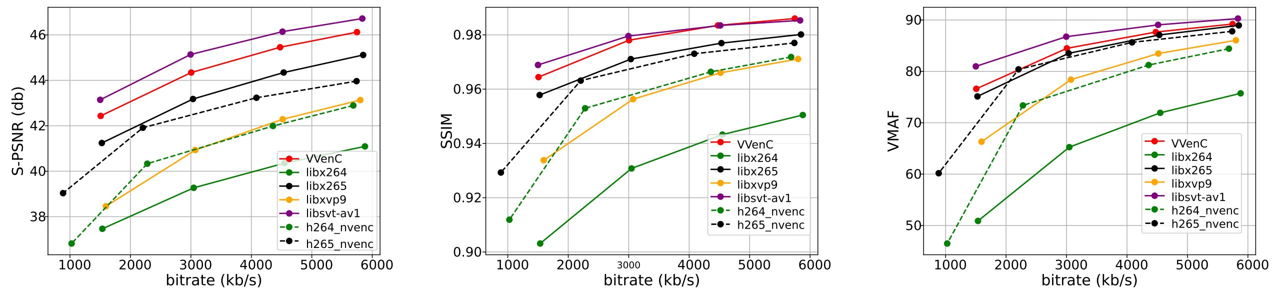
Standard	Software (version)	Hardware
AVC/H.264	libx264 [102] (v0.164.3106)	h264_nvenc [103]
HEVC/H.265	libx265 [104] (v3.5+1)	hevc_nvenc [105]
VVC/H.266	VVenC [106] (v1.7.0)	-
AV1	libsvtav1 [107] (v1.4.1)	-
VP9	libvpx-vp9 [108] (v1.11.0)	-

three distinct quality metrics: S-PSNR, SSIM, and VMAF. From the results, it is evident that the AV1 software encoder achieves the highest quality in terms of S-PSNR and VMAF across all four bitrates. The performance of VVenC software encoder is quite close to AV1, particularly at high bitrates, for the SSIM metric. On the other hand, the libx264 software encoder achieves the lowest quality among the tested encoders. It is worth noting that the hardware design for the AVC/H.264 standard significantly outperforms the libx264 software encoder across all quality metrics and bitrates. Interestingly, the software implementation of the HEVC/H.265 standard exhibits slightly higher quality than its hardware implementation. This can be attributed to increased focus on speed and complexity introduced by the new tools in the HEVC/H.265 standard, making the configurability of a hardware encoder for HEVC more challenging compared to the AVC/H.264 hardware encoders.

The associated BD-rate results concerning the AVC/H.264 software encoder for S-PSNR, SSIM, and VMAF are depicted in Figures 5(a), 5(b), and 5(c), respectively. These metrics are plotted against the encoding time. The results reveal that the hardware encoders (h264_nvenc and h265_nvenc) and the AV1 software encoder offer the best tradeoff between coding efficiency and encoding time. Notably, only the hardware encoders can achieve real-time encoding at 30 frames per second. To achieve real-time encoding, the AV1, AVC, and VP9 software encoders would require a powerful processor with multiple cores operating at a higher frequency. In contrast, the VVC/H.266 software encoder (VVenC) exhibits significantly longer encoding times, taking more than one hour to encode a 10-second video. The new coding tools introduced in the VVC/H.266 standard have expanded the search space for rate-distortion optimizations, leading to increased encoding complexity. To enable real-time capability, advanced algorithmic optimizations, along with more efficient low-level optimizations, are necessary. Furthermore, the development of efficient hardware designs for the VVC/H.266 standard becomes crucial for low-energy embedded devices to achieve real-time encoding and benefit from its high coding efficiency and advanced features for ODV contents.

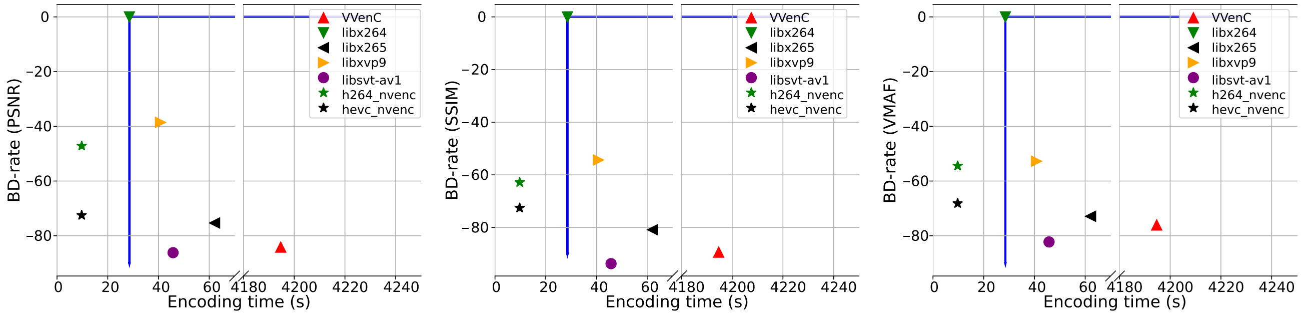
B. TESTBED for UAV 360° VIDEO STREAMING

The proposed testbed consists of a UAV equipped with a 360-degree camera and a 5G modem, and an edge server. The 360-degree camera captures a comprehensive view of the surroundings, providing an immersive 6DoF viewing experience. The 5G modem enables real-time transmission



(a) S-PSNR (b) SSIM (c) VMAF

FIGURE 4. The average quality in S-PSNR (dB), SSIM, and VMAF at different bit rates for the seven considered encoders.



(a) BD-PSNR (b) BD-SSIM (c) BD-VMAF

FIGURE 5. The BD-rate performance in S-PSNR (dB), SSIM, and VMAF versus encoding time for the seven considered encoders on 10-second video sequences.

of high-resolution footage from the UAV to the edge server. The user connects to the edge server through an HMD to view live 4K 360-degree video footage.

Figure 6(a) depicts the setup for the field tests conducted with a First Person View (FPV) UAV operator controlling the UAV in a desert environment. The operator sent commands to the UAV through a central server, located 100 km away from both the UAV and the operator. Both the UAV and the operator were connected to a consumer 5G network. The details of other settings are outlined in Table 12. During the experiment, the operator flew the drone at a fixed position, while varying the altitude. Simultaneously, the onboard computer of the UAV recorded the information received from the 5G modem, including the Cell ID, throughput, and network latency from the UAV to the central server.

Figures 7(a) and 7(b) provide insights into the handovers and the instantaneous throughput as a function of altitude in the scenario of vertical movement of the drone. In Figure 7(a), it can be observed that the drone experienced a total of ten handovers, across four available BSs that cover the flying area. Figure 7(b) shows that most handovers resulted in improved instantaneous throughput. However, the throughput exhibited significant fluctuations due to fluctuating wireless connectivity and interference. At higher altitudes, the drone encountered interference from BSs, primarily designed for ground-based users. Consequently, the latency and quality of the video and control signals degraded and thereby posed challenges for effective drone navigation by the operator. Our field tests showcased the control of UAV through 5G using a VR headset and 360-degree video feedback, at altitudes of up to 600 meters. These tests shed light on the potential challenges imposed by interfering BSs and suboptimal handover conditions in VR-based UAV control.

TABLE 12. Testbed Configuration for 360° Video Streaming over UAV.

Parameter	Value
5G Max(upload/download)	50 Mbps/100 Mbps
Server CPU	8 cores @ 2.5 GHz
Server memory	16 Gb
Distance UAV to Server	500 m
Distance VR HMD/UAV to Server	100 Km
UAV flight speed during tests	25Km/h
UAV's onboard computer	Jetson nano
UAV's weight	2.5 Kg
360-degree camera	Ricoh Theta Z1

IX. OPEN CHALLENGES

From Figure 7(b), it is evident that UAV communication, particularly at high altitudes and during mobility, is susceptible to significant throughput variation. This inherent issue raises concerns about attaining a high video quality and low G2G latency. To address these challenges, several open research directions need to be pursued. In the following, we describe a few prominent open directions.

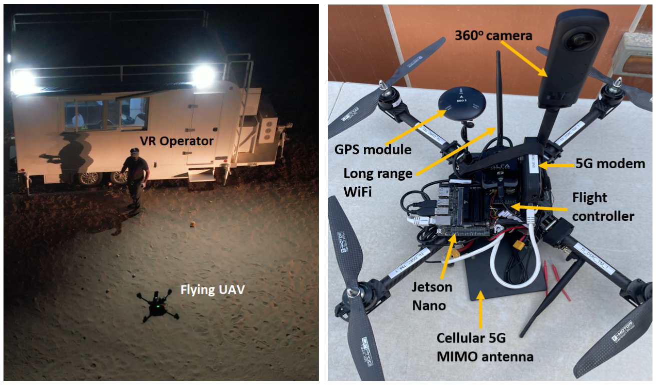
A. ADAPTIVE LOW-LATENCY 360° VIDEO STREAMING

Ensuring rapid and accurate adaptation of the video bitrate to channel throughput fluctuations is crucial to prevent buffering at both the transmitter and receiver, and thereby minimizing G2G latency. In this regard, utilizing information from the physical layer, as well as UAV status, position, and environmental conditions, can significantly enhance throughput prediction, and facilitate proactive adaptation of encoder parameters such as spatial resolution, temporal frame rate, quantization parameter, and projection format. Furthermore, sophisticated rate control mechanisms can further minimize G2G latency and maximize perceived quality. Advanced machine learning techniques, including deep reinforcement learning, have shown promise in bitrate adaptation while optimizing perceived video quality [109], [110]. However, leveraging these machine learning techniques for real-time bitrate adaptation remains an open challenge. Additionally, exploring advanced optimization techniques, like FoV prediction, can prioritize higher quality for the viewport of aerial ODVs, thereby improving bandwidth utilization and enhancing the user experience. Addressing these open research challenges will be pivotal in facilitating improved QoE, reduced latency, and superior video quality.

Further, as observed in Table 3, encoding complexity constitutes a major component of the G2G latency. Leveraging the latest video coding standards and efficient hardware encoders, such as hevc_nvenc, can substantially enhance perceived video quality. The hevc_nvenc encoder enables real-time encoding with low energy consumption [111], harnessing the coding efficiency promised by the advanced video coding standard, HEVC/H.265. This, in turn, extends the UAV's battery life. At the cloud level, more efficient software encoders like SVT-AV1 can be utilized for video transcoding. However, these standard codecs need to be benchmarked systematically by analyzing their encoding and decoding latency, quality, and error resilience [112] for a diverse range of receive Signal-to-noise-ratio (SNR) values and GOP sizes. In addition, neural-based codec designs can be explored to develop robust encoders to counter channel-induced errors [113] and enhance the quality.

B. COOPERATIVE AERIAL VIDEO STREAMING

Cooperative immersive video streaming, exemplified by Intel's Trueview [114], has the potential to enable a truly immersive viewing experience [115]. This approach allows users to independently select their preferred viewing angle by streaming from multiple cameras or sources, leveraging spatial diversity in terms of viewing angle, content, or geographic location. In multi-UAV applications, the individual UAVs collaboratively and cohesively capture videos, which are then synthesized into a panoramic video. Moreover, employing multiple UAVs enhances the immersive experience with 6DoF capabilities [116], [117]. However, developing a multi-UAV cooperative immersive video streaming system entails addressing a unique set of challenges in joint communication, computation, and control design. Cooperative aerial video streaming requires effective synchronization and coordination among the UAVs to ensure comprehensive scene coverage, without compromising QoE while minimizing network bandwidth usage. Additionally, capturing more dynamic events, such as sports or moving ground targets [115], [117], necessitates accurate motion prediction, such as player or target movement, which, in turn, relies on coordinated trajectory planning and 3D placement of UAVs, considering their battery levels in addition to the QoE.



(a) Field test

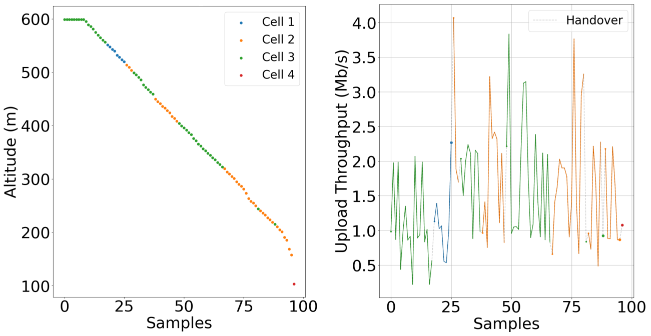
(b) UAV

FIGURE 6. Illustration of the field test setting and the UAV configuration.

Note that, streaming videos from all UAVs simultaneously poses a significant resource burden. To address this challenge, bandwidth-saving streaming techniques can be employed by leveraging users' attention information [118]. Specifically, UAVs whose videos are deemed unnoticed by users can remain idle during transmission. However, we argue that instead of staying idle, these UAVs can contribute to real-time video streaming, thus enhancing communication efficiency and throughput further. For instance, the UAV swarm can collectively form a virtual Multiple-input and Multiple-output (MIMO) system [119]. This type of MIMO system exhibits distinct wireless channel characteristics. Considering the unique channel model and the requirements for throughput and latency, designing a cooperative aerial video streaming for real-time and interactive panoramic videos poses considerable challenges. Addressing these challenges requires innovative solutions that account for coordination, resource optimization, wireless channel characteristics, throughput, and latency requirements.

C. QoE-AWARE CONTROL AND COMMUNICATION

In this section, we discuss mechanisms to support the high data rates and low latency required for real-time transmission of aerial ODVs [120]. UAV-based real-time ODV streaming represents a distinct class of services, encompassing both



(a) Handover vs. altitude. (b) Throughput vs. handover.

FIGURE 7. Handovers and instantaneous throughput performance versus the drone altitude in vertical landing flying conditions. The average throughput values of the cells in green, orange, black, and red are 14.55 Mbps, 17.19 Mbps, 11.21 Mbps, and 10.79 Mbps, respectively.

enhanced mobile broadband and ultra-reliable low-latency traffic, necessitating novel communication designs. Additional challenges arise due to their dynamic topology and limited energy resources [121], [122], requiring judicious resource allocation strategies [123]. The channel quality and network throughput of aerial users are also influenced by their flight trajectory, necessitating the orchestration of joint QoE-aware resource allocation and drone route selection mechanisms [124], [125].

One potential approach is to develop QoE-aware Network Slicing (NS) mechanisms. Unlike traditional QoS-based NS [126], a dynamic NS framework is needed that considers UAV mobility and position, optimizes energy levels, and ensures minimal resource overhead. The NS scheme must also guarantee strong isolation to minimize the impact on ground-based users. In multi-UAV streaming systems for 360° videos [65], additional challenges arise in resource allocation among UAVs. Each UAV can independently adjust its encoding bitrate and position [127], competing for resources with other UAVs in the swarm.

In addition, the design of schemes leveraging video saliency to predict users' FoV and employing multicast transmission techniques based on users' locations and FoV correlations can be studied, as grouping and multicasting can improve network throughput and QoE [128], [129]. Additionally, the design of policies adapting the encoding bitrate of tiles based on channel quality, available resources, and content quality, can further enhance QoE [120]. Furthermore, in applications involving the teleoperation of UAVs, such as fire disaster monitoring [82] and suspicious vehicle tracking [130], the QoS relies on the interplay between control command delivery and video data transmission. The latency experienced in one link can impact the latency budget of the other link. Moreover, unreliable control command communication can influence the UAV's reaction and view angle, resulting in undesired information for the remote operator. Therefore, the entanglement and mutual influence between control command delivery and real-time aerial video transmission require dedicated consideration in the design.

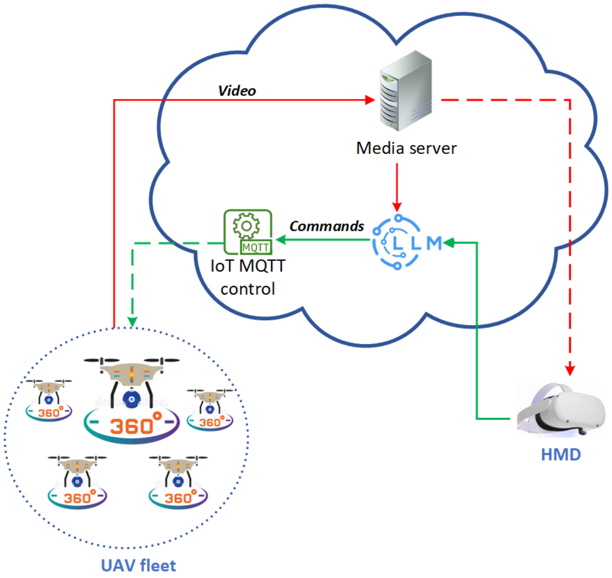


FIGURE 8. Use case scenario for LLMs control commands for UAV with 360° camera.

D. DESIGN OF COMMUNICATION PROTOCOLS TAILORED FOR UAV-BASED VIDEO STREAMING

Transmission Control Protocol (TCP), due to its inherent limitations such as connection delays and head-of-line blocking, poses challenges in delivering satisfactory QoE for real-time 360° video streaming. Moreover, its implementation within the operating system kernel hinders the development and deployment of variants that can be optimized using application-layer data (e.g., FoV) and other parameters like UAV position [131]. To address these limitations, protocols like QUIC [132] have been proposed. Notably, Park *et al.* [131] introduced a cross-layer scheduling mechanism for QUIC, leveraging both application-layer data (e.g., object sizes and priorities) and network-layer information. Such tailored designs, incorporating specific characteristics of video streaming and the unique attributes of A2G channels, hold significant potential for enhancing performance.

Conversely, adopting a semantic communication approach [133], where the emphasis is on effectively conveying the intended meaning of information rather than merely transmitting raw data, holds promise for enabling various applications reliant on real-time streaming from UAV platforms. However, to leverage the benefits of joint optimization using both video content and physical layer data, the development of customizable communication protocols is crucial. For example, similar to [134], utilizing bandwidth estimation provided by the WebRTC protocol can enable optimization of encoding parameters at the application layer, thereby impacting frame drops and latency.

Furthermore, reported testbed studies frequently rely on existing protocols without tailored optimization for video streaming, leading to an inaccurate assessment of real-world performance. A comprehensive evaluation utilizing wireless protocols specifically optimized for UAV-based real-time 360° video streaming is crucial to reveal the true state-of-the-art performance achievable in practical scenarios.

E. LLM FOR IMMERSIVE VIDEO STREAMING

The rapid advancement in natural language processing has paved the way for the development of Large Language Model (LLM) like BERT [135], GPT-3/GPT-4, and FALCON. These versatile models push the state-of-the-art on many downstream tasks, finding applications in various domains, including conversation, medicine, telecommunications [136], and robotics [137]. In the context of streaming 360-degree video from one or multiple UAVs, illustrated in Fig. 8, leveraging these LLMs can greatly enhance performance. In the following, we describe some examples to illustrate the potential of LLMs in enhancing the performance of real-time streaming of aerial 360° videos.

In control scenarios, end-users can provide task prompts to the LLM along with descriptions of the environment captured by the 360-degree camera. The LLM can then generate commands for the UAVs to successfully execute tasks while minimizing energy consumption and avoiding obstacles. Notably, the description of the surrounding environment can be provided either by the end-user or automatically generated by leveraging vision-language models, such as SimVLM [138], Flamingo [139], or BLIP-2 [140].

Other use cases integrate LLM and Multimodal Large Language Model (MLLM) into the video streaming framework for enhanced compression efficiency. The first use case involves the application of LLM for the lossless compression of images or videos, serving as an entropy encoder. Recent research, from DeepMind [141], underscores the potent versatility of LLMs as general-purpose compressors, owing to their in-context learning capabilities. Experiments utilizing Chinchila 70B, solely trained in natural language, revealed impressive compression ratios, achieving 43.4% on ImageNet patches. Notably, this rate outperforms domain-specific image compressors such as Portable Network Graphics (PNG) (58.5%). The second use case harnesses MLLM shared at both the transmitter and receiver for a lossy coding setting. The transmitter first generates an accurate description of the image or video content through the image captioning capability of the MLLM. Instead of transmitting the image or video, the text description (semantic information) is then sent to the receiver, requiring a significantly lower data rate. At the receiver, the generative capability of the MLLM is used to reconstruct the image or video based on the received text description. In addition to the text prompt, the generation can also be guided by side information like edge map, color map to generate a more compelling representation. In the third use case, the MLLM is employed solely at the transmitter to leverage its code-generation capability, representing the image or video for transmission. Subsequently, the code, requiring a lower data rate, is shared with the receiver, enabling direct utilization to render the image or video through the code description [142].

The above examples illustrate the tantalizing potential of LLMs in not only improving the compression performance but also in configuring the physical layer parameters [143]. Nonetheless, the LLMs still suffer from long inference time and high memory requirement which needs to be addressed to leverage LLMs for enhancing the performance of real-time streaming of aerial ODVs. In addition, there is a need for development of customized LLMs.

X. CONCLUSION

In this paper, we conducted a thorough survey of challenges and research efforts in UAV-based immersive video streaming. By enabling immersive viewing with up to 6DoF, this technology enhances the QoE for various applications such as surveillance, autonomous driving, healthcare, and education. However, real-time streaming of aerial 360-degree videos poses unique challenges in terms of communications, computation, and control, owing to the unique characteristics of the UAV-to-ground wireless channel and limited onboard energy availability. We highlighted these challenges by first reviewing the key components of 360-degree video streaming over A2G wireless channels and reviewed the technology used to achieve low end-to-end latency. Additionally, we introduced a new dataset consisting of ten 360-degree videos captured by UAV in diverse flying conditions, enabling us to evaluate the coding efficiency and complexity of various software and hardware video encoders. Through our experiments, we found that only hardware implementations of AVC/H.264 and HEVC/H.265 encoders achieved real-time encoding, making them suitable for UAV platforms, with limited computing and energy resources. Furthermore, the AV1 encoder demonstrated the best coding performance, albeit with high complexity, and therefore can be utilized for efficient video transcoding on more powerful devices in the cloud. Moreover, we presented a testbed for 360-degree video streaming over a drone with 5G communication, illustrating the impact of mobility on interference, handovers, and video quality. Finally, we discussed open challenges and future research directions to enhance the key performance metrics of live immersive video streaming over UAVs.

This paper delves into real-time streaming of omnidirectional videos captured via UAVs, offering valuable insights to enhance the QoE in this domain. The findings in the paper pave the way for further advancements in live immersive UAV video streaming, ultimately benefiting a broad spectrum of applications and industries.

REFERENCES

[1] J. van der Hooft, H. Amirpour, M. T. Vega, Y. Sanchez, R. Schatz, T. Schierl, C. Timmerer, A tutorial on immersive video delivery: From omnidirectional video to holography, IEEE Communications Surveys & Tutorials 25 (2) (2023) 1336–1375. doi:10.1109/COMST.2023.3263252.

[2] A. Bentaleb, M. Lim, M. N. Akcay, A. C. Begen, S. Hammoudi, R. Zimmermann, Toward one-second latency: Evolution of live media

streaming (2023). arXiv:2310.03256.
URL https://arxiv.org/abs/2310.03256

[3] 3GPP TR 26.925, Typical traffic characteristics of media services on 3gpp networks (rel. 18)Accessed: 1-July-2024.
URL https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3533

[4] T. Izydorczyk, G. Berardinelli, P. Mogensen, M. M. Ginard, J. Wigard, I. Z. Kovács, Achieving high uav uplink throughput by using beamforming on board, IEEE Access 8 (2020) 82528–82538. doi:10.1109/ACCESS.2020.2991658.

[5] A. Zaki-Hindi, R. Amorim, I. Z. Kovács, J. Wigard, Uplink coexistence for high throughput uavs in cellular networks, in: GLOBECOM 2022 - 2022 IEEE Global Communications Conference, 2022, pp. 2957–2962. doi:10.1109/GLOBECOM48099.2022.10001239.

[6] T.-V. Nguyen, N. P. Nguyen, C. Kim, N.-N. Dao, Intelligent aerial video streaming: Achievements and challenges, J. Netw. Comput. Appl. 211 (2023) 103564. doi:https://doi.org/10.1016/j.jnca.2022.103564.

[7] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, M. Debbah, A tutorial on UAVs for wireless networks: Applications, challenges, and open problems, IEEE Commun. Surveys Tuts. 21 (3) (2019) 2334–2360. doi:10.1109/COMST.2019.2902862.

[8] Q. Wu, J. Xu, Y. Zeng, D. W. K. Ng, N. Al-Dhahir, R. Schober, A. L. Swindlehurst, A comprehensive overview on 5G-and-beyond networks with UAVs: From communications to sensing and intelligence, IEEE J. Sel. Areas Commun. 39 (10) (2021) 2912–2945. doi:10.1109/JSAC.2021.3088681.

[9] S. Hayat, E. Yanmaz, R. Muzaffar, Survey on unmanned aerial vehicle networks for civil applications: A communications viewpoint, IEEE Commun. Surveys Tuts. 18 (4) (2016) 2624–2661. doi:10.1109/COMST.2016.2560343.

[10] A. Baltaci, E. Dinc, M. Ozger, A. Alabbasi, C. Cavdar, D. Schupke, A survey of wireless networks for future aerial communications (FACOM), IEEE Commun. Surveys Tuts. 23 (4) (2021) 2833–2884. doi:10.1109/COMST.2021.3103044.

[11] D. Mishra, A. M. Vegni, V. Loscrí, E. Natalizio, Drone networking in the 6G era: A technology overview, IEEE Commun. Standards Mag. 5 (4) (2021) 88–95. doi:10.1109/MCOMSTD.0001.2100016.

[12] C. Yan, L. Fu, J. Zhang, J. Wang, A comprehensive survey on UAV communication channel modeling, IEEE Access 7 (2019) 107769–107792. doi:10.1109/ACCESS.2019.2933173.

[13] Z. Xiao, L. Zhu, Y. Liu, P. Yi, R. Zhang, X.-G. Xia, R. Schober, A survey on millimeter-wave beamforming enabled UAV communications and networking, IEEE Commun. Surveys Tuts. 24 (1) (2022) 557–610. doi:10.1109/COMST.2021.3124512.

[14] Y. Zeng, Q. Wu, R. Zhang, Accessing from the sky: A tutorial on UAV communications for 5G and beyond, Proc. IEEE 107 (12) (2019) 2327–2375. doi:10.1109/JPROC.2019.2952892.

[15] T. Elmokadem, A. V. Savkin, Towards fully autonomous UAVs: A survey, Sensors 21 (18) (2021) 1–39. doi:10.3390/s21186223.

[16] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia-Rodriguez, J. Yuan, Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges, IEEE Commun. Surveys Tuts. 21 (4) (2019) 3417–3442. doi:10.1109/COMST.2019.2906228.

[17] V. Marojevic, I. Guvenc, R. Dutta, M. L. Sichitiu, B. A. Floyd, Advanced wireless for unmanned aerial systems: 5G standardization, research challenges, and AERPAW architecture, IEEE Veh. Technol. Mag. 15 (2) (2020) 22–30. doi:10.1109/MVT.2020.2979494.

[18] A. S. Abdalla, V. Marojevic, Communications standards for unmanned aircraft systems: The 3GPP perspective and research drivers, IEEE Commun. Standards Mag. 5 (1) (2021) 70–77. doi:10.1109/MCOMSTD.001.2000032.

[19] H. Wang, H. Zhao, J. Zhang, D. Ma, J. Li, J. Wei, Survey on unmanned aerial vehicle networks: A cyber physical system perspective, IEEE Commun. Surveys Tuts. 22 (2) (2020) 1027–1070. doi:10.1109/COMST.2019.2962207.

[20] A. Yaqoob, T. Bi, G.-M. Muntean, A survey on adaptive 360° video streaming: Solutions, challenges and opportunities, IEEE Commun. Surveys Tuts.s 22 (4) (2020) 2801–2838. doi:10.1109/COMST.2020.3006999.

[21] D. He, C. Westphal, J. J. Garcia-Luna-Aceves, Network support for AR/VR and immersive video application: A survey, in: Proc. 15th Int. Joint Conf. e-Business Telecommun., 2018, pp. 359–369.

[22] M. Zink, R. Sitaraman, K. Nahrstedt, Scalable 360° video stream delivery: Challenges, solutions, and opportunities, Proc. IEEE 107 (4) (2019) 639–650. doi:10.1109/JPROC.2019.2894817.

[23] C. Bonnineau, W. Hamidouche, J. Fournier, N. Sidaty, J.-F. Travers, O. Déforges, Perceptual quality assessment of HEVC and VVC standards for 8K video, IEEE Trans. Broadcast. 68 (1) (2022) 246–253.

[24] I.-C. Lo, K.-T. Shih, H. H. Chen, Efficient and accurate stitching for 360° dual-fisheye images and videos, IEEE Trans. Image Process. 31 (2022) 251–262. doi:10.1109/TIP.2021.3130531.

[25] Y. Ye, J. M. Boyce, P. Hanhart, Omnidirectional 360° video coding technology in responses to the joint call for proposals on video compression with capability beyond HEVC, IEEE Transactions on Circuits and Systems for Video Technology 30 (5) (2020) 1241–1252. doi:10.1109/TCSVT.2019.2953827.

[26] E. Kuzyakov, D. Pio, Next-generation video encoding techniques for 360 video and VR, accessed on 1-July-2023 (Jan. 2016).
URL https://engineering.fb.com/2016/01/21/virtual-reality/next-generation-video-encoding-techniques-for-360-video-and-vr/

[27] C. Zhou, Z. Li, Y. Liu, A measurement study of oculus 360 degree video streaming, in: Proceedings of the 8th ACM on Multimedia Systems Conference, MMSys'17, Association for Computing Machinery, New York, NY, USA, 2017, p. 27–37. doi:10.1145/3083187.3083190.

[28] S. Wang, X. Zhang, M. Xiao, K. Chiu, Y. Liu, SphericRTC: A system for content-adaptive real-time 360-degree video communication, in: Proceedings of the 28th ACM International Conference on Multimedia, MM '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 3595–3603. doi:10.1145/3394171.3413999.

[29] T. Wiegand, G. Sullivan, G. Bjontegaard, A. Luthra, Overview of the H.264/AVC video coding standard, IEEE Trans. Circuits Syst. Video Technol. 13 (7) (2003) 560–576. doi:10.1109/TCSVT.2003.815165.

[30] G. J. Sullivan, J.-R. Ohm, W.-J. Han, T. Wiegand, Overview of the high efficiency video coding (HEVC) standard, IEEE Trans. Circuits Syst. Video Technol. 22 (12) (2012) 1649–1668. doi:10.1109/TCSVT.2012.2221191.

[31] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, J.-R. Ohm, Overview of the versatile video coding (VVC) standard and its applications, IEEE Trans. Circuits Syst. Video Technol. 31 (10) (2021) 3736–3764. doi:10.1109/TCSVT.2021.3101953.

[32] K. Misra, A. Segall, M. Horowitz, S. Xu, A. Fuldseth, M. Zhou, An overview of tiles in HEVC, IEEE J. Sel. Topics Signal Process. 7 (6) (2013) 969–977. doi:10.1109/JSTSP.2013.2271451.

[33] J. Son, D. Jang, E.-S. Ryu, Implementing motion-constrained tile and viewport extraction for VR streaming, in: Proc. 28th ACM SIGMM Workshop Netw. Oper. Syst. Support Digit. Audio Video, 2018, pp. 61–66.

[34] N. Mahmoudian Bidgoli, T. Maugey, A. Roumy, Fine granularity access in interactive compression of 360-degree images based on rate-adaptive channel codes, IEEE Trans. Multimedia 23 (2021) 2868–2882. doi:10.1109/TMM.2020.3017890.

[35] Y. Zhou, L. Tian, C. Zhu, X. Jin, Y. Sun, Video coding optimization for virtual reality 360-degree source, IEEE J. Sel. Topics Signal Process. 14 (1) (2020) 118–129. doi:10.1109/JSTSP.2019.2957952.

[36] Y.-C. Su, K. Grauman, Learning compressible 360° video isomers, IEEE Trans. Pattern Anal. Mach. Intell. 43 (8) (2021) 2697–2709. doi:10.1109/TPAMI.2020.2974472.

[37] D. Minnen, J. Ballé, G. D. Toderici, Joint autoregressive and hierarchical priors for learned image compression, Adv. Neural Inf. Process. Syst. 31 (2018).

[38] J. Cao, X. Su, B. Finley, A. Pauanne, M. Ammar, P. Hui, Evaluating multimedia protocols on 5G edge for mobile augmented reality, in: Proc. 17th Int. Conf. Mobility, Sens. Netw., 2021, pp. 199–206. doi:10.1109/MSN53354.2021.00042.

[39] M. M. Hannuksela, Y.-K. Wang, An overview of omnidirectional media format (OMAF), Proc. IEEE 109 (9) (2021) 1590–1606. doi:10.1109/JPROC.2021.3063544.

[40] B. Sredojev, D. Samardzija, D. Posarac, WebRTC technology overview and signaling solution design and implementation, in:

Proc. 38th Int. Convention Inf. Commun. Technol., Electron. Microelectron., 2015, pp. 1006–1009. doi:10.1109/MIPRO.2015.7160422.

[41] K. K. Sreedhar, I. D. D. Curcio, A. Hourunranta, M. Lepistö, Immersive media experience with MPEG OMAF multi-viewpoints and overlays, in: Proc. 11th ACM Multimedia Syst. Conf., 2020, pp. 333–336.

[42] TileMedia, How ClearVR Drives and Leverages Standards, accessed on 1-July-2023 (Jan. 2022).
URL https://www.tiledmedia.com/how-clearvr-drives-and-leverages-standards/

[43] N. Sehad, B. Cherif, I. Khadraoui, W. Hamidouche, F. Bader, R. Jäntti, M. Debbah, Locomotion-based UAV control toward the internet of senses, IEEE Transactions on Circuits and Systems II: Express Briefs 70 (5) (2023) 1804–1808. doi:10.1109/TCSII.2023.3257363.

[44] Intel, Immersive video sample reference implementation, accessed on 1-July-2023 (Dec. 2022).
URL https://www.intel.com/content/www/us/en/developer/articles/technical/immersive-video-sample-powered-by-ovc.html

[45] L. Wang, X. Shi, Y. Liu, Foveated rendering: A state-of-the-art survey, Computational Visual Media 9 (2) (2023) 195–228. doi:10.1007/s41095-022-0306-4.
URL https://doi.org/10.1007/s41095-022-0306-4

[46] W. Cheung, Gang Shen and Dusty Robbins, Solution implementation summary: Media 360-degree video distributionAccessed: 19-january-2023.
URL https://www.intel.it/content/dam/www/central-libraries/us/en/documents/advanced-360video-implementation-summary-final.pdf

[47] M. Warburton, M. Mon-Williams, F. Mushtaq, J. R. Morehead, Measuring motion-to-photon latency for sensorimotor experiments with virtual reality systems, Springer - Behavior Research Methods 10 (Oct. 2022).

[48] H. T. T. Tran, N. P. Ngoc, C. M. Bui, M. H. Pham, T. C. Thang, An evaluation of quality metrics for 360 videos, in: 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN), 2017, pp. 7–11. doi:10.1109/ICUFN.2017.7993736.

[49] A. Katsenou, J. Mao, I. Mavromatis, Energy-rate-quality tradeoffs of state-of-the-art video codecs (2022).
URL https://europepmc.org/article/PPR/PPR556327

[50] J. Wubben, F. Fabra, C. Calafate, T. Krzeszowski, J. Marquez-Barja, J.-C. Cano, P. Manzoni, Accurate landing of unmanned aerial vehicles using ground pattern recognition, Electronics 8 (2019) 1532. doi:10.3390/electronics8121532.

[51] U. Veyna, S. Garcia-Nieto, R. Simarro, J. V. Salcedo, Quadcopters testing platform for educational environments, Sensors 21 (12) (2021). doi:10.3390/s21124134.
URL https://www.mdpi.com/1424-8220/21/12/4134

[52] N. Gao, Y. Zeng, J. Wang, D. Wu, C. Zhang, Q. Song, J. Qian, S. Jin, Energy model for UAV communications: Experimental validation and model generalization, China Commun. 18 (7) (2021) 253–264.

[53] H. Yan, S.-H. Yang, Y. Ding, Y. Chen, Energy consumption models for UAV communications: A brief survey, in: Proc. IEEE Int. Conf. Internet Things and IEEE Green Comput. Commun. and IEEE Cyber, Physical Social Comput. and IEEE Smart Data and IEEE Congr. Cybermatics, 2022, pp. 161–167.

[54] C. You, R. Zhang, 3D trajectory optimization in Rician fading for UAV-enabled data harvesting, IEEE Transactions on Wireless Communications 18 (6) (2019) 3192–3207. doi:10.1109/TWC.2019.2911939.

[55] C. You, R. Zhang, Hybrid offline-online design for UAV-enabled data harvesting in probabilistic LoS channels, IEEE Trans. Wireless Commun. 19 (6) (2020) 3753–3768.

[56] A. Al-Hourani, S. Kandeepan, S. Lardner, Optimal LAP altitude for maximum coverage, IEEE Wireless Commun. Lett. 3 (6) (2014) 569–572.

[57] H. V. Abeywickrama, B. A. Jayawickrama, Y. He, E. Dutkiewicz, Empirical power consumption model for uavs, in: 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), 2018, pp. 1–5. doi:10.1109/VTCFall.2018.8690666.

[58] V. V. Menon, C. Feldmann, K. Schoeffmann, M. Ghanbari, C. Timmerer, Green video complexity analysis for efficient encoding in adaptive video streaming, in: Proceedings of the First International Workshop on Green Multimedia Systems, GMSys ’23, ACM, 2023. doi:10.1145/3593908.3593942.
URL http://dx.doi.org/10.1145/3593908.3593942

[59] S. M. Satti, M. Obermann, C. Schmidmer, M. Keyhl, Video encoding complexity characterization, in: 2022 14th International Conference on Quality of Multimedia Experience (QoMEX), 2022, pp. 1–4. doi:10.1109/QoMEX55416.2022.9900917.

[60] V. I. Forum, VR Industry Forum Guidelines V2.3, accessed on 1-July-2023 (Jan. 2021).
URL https://www.vr-if.org/wp-content/uploads/vrif2020.180.00-Guidelines-2.3_clean..pdf

[61] F. Qian, L. Ji, B. Han, V. Gopalakrishnan, Optimizing 360 video delivery over cellular networks, in: Proc. 5th Workshop All Things Cellular: Oper., Appl. Challenges, 2016, pp. 1–6.

[62] TileMedia, Tiled Media High Quality VR streaming, accessed on 1-July-2023 (Jan. 2022).
URL https://www.tiledmedia.com/high-quality-vr-streaming/

[63] X. Xiao, W. Wang, T. Chen, Y. Cao, T. Jiang, Q. Zhang, Sensor-augmented neural adaptive bitrate video streaming on UAVs, IEEE Trans. Multimedia 22 (6) (2020) 1567–1576.

[64] R. Muzaffar, E. Yanmaz, C. Raffelsberger, C. Bettstetter, A. Cavallaro, Live multicast video streaming from drones: An experimental study, Auton. Robot. 44 (2020) 75–91.

[65] C. He, Z. Xie, C. Tian, A QoE-oriented uplink allocation for multi-UAV video streaming, Sensors 19 (15) (2019) 1–19.

[66] Q. Zhang, J. Miao, Z. Zhang, F. R. Yu, F. Fu, T. Wu, Energy-efficient video streaming in UAV-enabled wireless networks: A safe-DQN approach, in: Proc. IEEE Global Commun. Conf., 2020, pp. 1–7.

[67] L. Zhang, J. Chakareski, UAV-assisted edge computing and streaming for wireless virtual reality: Analysis, algorithm design, and performance guarantees, IEEE Trans. Veh. Technol. 71 (3) (2022) 3267–3275.

[68] F. Hu, Y. Deng, A. H. Aghvami, Cooperative multigroup broadcast 360° video delivery network: A hierarchical federated deep reinforcement learning approach, IEEE Trans. Wireless Commun. 21 (6) (2022) 4009–4024.

[69] Y. Guo, Y. Chen, J. Hu, H. Zheng, Trajectory planning of UAV with real-time video transmission based on genetic algorithm, in: Proc. Cross Strait Radio Sci. Wireless Technol. Conf., 2020, pp. 1–3.

[70] C. Zhan, H. Hu, Z. Wang, R. Fan, D. Niyato, Unmanned aircraft system aided adaptive video streaming: A joint optimization approach, IEEE Trans. Multimedia 22 (3) (2020) 795–807.

[71] L. A. b. Burhanuddin, X. Liu, Y. Deng, U. Challita, A. Zahemszky, QoE optimization for live video streaming in UAV-to-UAV communications via deep reinforcement learning, IEEE Trans. Veh. Technol. 71 (5) (2022) 5358–5370.

[72] J. Chakareski, UAV-IoT for next generation virtual reality, IEEE Trans. Image Process. 28 (12) (2019) 5977–5990.

[73] M. Khan, J. Chakareski, S. Gupta, RF-FSO dual-path UAV network for high fidelity multi-viewpoint scalable 360° video streaming, in: Proc. IEEE 22nd Int. Workshop Multimedia Signal Process., 2020, pp. 1–6.

[74] A. Stornig, A. Fakhreddine, H. Hellwagner, P. Popovski, C. Bettstetter, Video quality and latency for UAV teleoperation over LTE: A study with ns3, in: Proc. IEEE 93rd Veh. Technol. Conf., 2021, pp. 1–7.

[75] H. Zhou, F. Hu, M. Juras, A. B. Mehta, Y. Deng, Real-time video streaming and control of cellular-connected UAV system: Prototype and performance evaluation, IEEE Wireless Commun. Lett. 10 (8) (2021) 1657–1661.

[76] J. Jin, J. Ma, L. Liu, L. Lu, G. Wu, D. Huang, N. Qin, Design of UAV video and control signal real-time transmission system based on 5G network, in: Proc. IEEE 16th Conf. Ind. Electron. Appl., 2021, pp. 533–537.

[77] T. Taleb, N. Sehad, Z. Nadir, J. Song, VR-based immersive service management in b5g mobile systems: A uav command and control use case, IEEE Internet of Things Journal 10 (6) (2023) 5349–5363. doi:10.1109/JIOT.2022.3222282.

[78] S. Qazi, A. S. Siddiqui, A. I. Wagan, UAV based real time video surveillance over 4G LTE, in: Proc. Int. Conf. Open Source Syst. Technol., 2015, pp. 141–145.

[79] C. Singhal, B. N. Chandana, Aerial-SON: UAV-based self-organizing network for video streaming in dense urban scenario, in: Proc. Int. Conf. Commun. Syst. Netw., 2021, pp. 7–12.

[80] M. Naveed, S. Qazi, B. A. Khawaja, M. Mustaqim, Evaluation of video streaming capacity of UAVs with respect to channel variation in 4G-LTE surveillance architecture, in: Proc. 8th Int. Conf. Inf. Commun. Technol., 2019, pp. 149–154.

[81] Z. Liu, Y. Jiang, Cross-layer design for UAV-based streaming media transmission, IEEE Trans. Circuits Syst. Video Technol. 32 (7) (2022) 4710–4723.

[82] K. Nihei, N. Kai, Y. Maruyama, T. Yamashita, D. Kanetomo, T. Kitahara, M. Maruyama, T. Ohki, K. Kusin, H. Segah, Forest fire surveillance using live video streaming from UAV via multiple LTE networks, in: Proc. IEEE 19th Annu. Consum. Commun. Netw. Conf., 2022, pp. 465–468.

[83] T. Yu, Y. Takaku, Y. Kaieda, K. Sakaguchi, Design and PoC implementation of mmWave-based offloading-enabled UAV surveillance system, IEEE Open J. Veh. Technol. 2 (2021) 436–447.

[84] X. Lin, V. Yajnanarayana, S. D. Muruganathan, S. Gao, H. Asplund, H.-L. Määttänen, M. Bergstrom, S. Euler, Y.-P. E. Wang, The sky is not the limit: LTE for unmanned aerial vehicles, IEEE Commun. Mag. 56 (4) (2018) 204–210.

[85] 3GPP TR 36.777, Enhanced LTE support for aerial vehicles.
URL https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3231

[86] 3GPP TS 22.125, Unmanned aerial system (UAS) support in 3GPPAccessed: 1-July-2023.
URL https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3545

[87] ATIS-I-0000092, 3GPP Release 17 – building blocks for uav applicationsAccessed: 1-July-2023.
URL https://access.atis.org/apps/group_public/download.php/66824/ATIS-I-0000092.pdf

[88] 3GPP TS 26.918, Virtual reality (VR) media services over 3gppAccessed: 1-July-2023.
URL https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3053

[89] 3GPP TS 26.501, 5G media streaming (5GMS); general description and architectureAccessed: 1-July-2023.
URL https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3582

[90] 3GPP TS 26.506, 5G real-time media communication architectureAccessed: 1-July-2023.
URL https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=4102

[91] 3GPP TS 26.522, 5G real-time media transport protocol configurationsAccessed: 1-July-2024.
URL https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=4114

[92] 3GPP TS 26.803, Study on 5G media streaming extensions for edge processingAccessed: 1-July-2023.
URL https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3742

[93] 3GPP TR 26.927, Study on Artificial Intelligence and Machine learning in 5G media servicesAccessed: 1-July-2023.
URL https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=4040

[94] A.-F. Perrin, V. Krassanakis, L. Zhang, V. Ricordel, M. Perreira Da Silva, O. Le Meur, EyeTrackUAV2: A large-scale binocular eye-tracking dataset for UAV videos, Drones 4 (1) (2020).

[95] K. Fu, J. Li, Y. Zhang, H. Shen, Y. Tian, Model-guided multi-path knowledge aggregation for aerial saliency prediction, IEEE Trans. Image Process. 29 (2020) 7117–7127. doi:10.1109/TIP.2020.2998977.

[96] S. Colonnese, F. Cuomo, L. Ferranti, T. Melodia, Efficient video streaming of 360° cameras in unmanned aerial vehicles: An analysis of real video sources, in: Proc. 7th Eur. Workshop Visual Inf. Process., 2018, pp. 1–6.

[97] T.-W. Mi, M.-T. Yang, Comparison of tracking techniques on 360-degree videos, Appl. Sci. 9 (16) (2019) 3336.

[98] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, P. L. Callet, A dataset of head and eye movements for 360° videos, in: Proc. 9th ACM Multimedia Syst. Conf., 2018, pp. 432–437.

[99] A. T. Nasrabadi, A. Samiei, A. Mahzari, R. P. McMahan, R. Prakash, M. C. Q. Farias, M. M. Carvalho, A taxonomy and dataset for 360° videos, in: Proc. 10th ACM Multimedia Syst. Conf., 2019, pp. 273–278.

[100] C. Wu, Z. Tan, Z. Wang, S. Yang, A dataset for exploring user behaviors in VR spherical video streaming, in: Proc. 8th ACM Multimedia Syst. Conf., 2017, pp. 193–198.

[101] 3GPP TR 26.925, Typical traffic characteristics of media services on 3GPP networks (rel. 16).
URL https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3533

[102] VideoLAN, libx264, https://www.videolan.org/developers/x264.html, accessed: 1-July-2023 (2013).

[103] NVIDIA Corporation, h264_nvenc, https://developer.nvidia.com/nvidia-video-codec-sdk, accessed: 1-July-2023 (2018).

[104] MulticoreWare Inc., x265, https://www.videolan.org/developers/x265.html, version: 3.5 (2021).

[105] NVIDIA Corporation, hevc_nvenc, https://developer.nvidia.com/nvidia-video-codec-sdk, sDK Version: 11.0 (May 2021).

[106] A. Wieckowski, J. Brandenburg, T. Hinz, C. Bartnik, V. George, G. Hege, C. Helmrich, A. Henkel, C. Lehmann, C. Stoffers, I. Zupancic, B. Bross, D. Marpe, Vvenc: An open and optimized vvc encoder implementation, in: 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2021, pp. 1–2. doi:10.1109/ICMEW53276.2021.9455944.

[107] Open Visual Cloud, Svt-av1, https://www.openvisualcloud.org/, version: 0.8.7 (2021).

[108] WebM Project, libvpx, https://www.webmproject.org/code/, version: 1.10.0 (2021).

[109] T. Huang, R.-X. Zhang, L. Sun, Deep reinforced bitrate ladders for adaptive video streaming, in: Proceedings of the 31st ACM Workshop on Network and Operating Systems Support for Digital Audio and Video, 2021, pp. 66–73.

[110] A. Mandhane, A. Zhernov, M. Rauh, C. Gu, M. Wang, F. Xue, W. Shang, D. Pang, R. Claus, C.-H. Chiang, et al., Muzero with self-competition for rate control in vp9 video compression, arXiv preprint arXiv:2202.06626 (2022).

[111] I. Farhat, I. Khadraoui, W. Hamidouche, M. K. Sharma, Energy cost of coding omnidirectional videos using arm and x86 platforms, in: Proceedings of the Second International ACM Green Multimedia Systems Workshop, GMSys '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 8–12. doi:10.1145/3652104.3652526.
URL https://doi.org/10.1145/3652104.3652526

[112] M. K. Sharma, I. Farhat, W. Hamidouche, Error concealment capacity analysis of standard video encoders for real-time immersive video streaming from UAVs, in: 2024 IEEE Wireless Communications and Networking Conference (WCNC), 2024, pp. 1–6. doi:10.1109/WCNC57260.2024.10570883.

[113] Y. Cheng, Z. Zhang, H. Li, A. Arapin, Y. Zhang, Q. Zhang, Y. Liu, X. Zhang, F. Y. Yan, A. Mazumdar, N. Feamster, J. Jiang, Grace: Loss-resilient real-time video through neural codecs (2024). arXiv:2305.12333.

[114] Intel, True view intel sports, accessed: 1-July-2023 (2021).
URL https://www.intel.com/content/www/us/en/sports/technology/true-view.html

[115] X. Wang, A. Chowdhery, M. Chiang, Networked drone cameras for sports streaming, in: Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst., 2017, pp. 308–318.

[116] X. Meng, W. Wang, B. Leong, SkyStitch: A cooperative multi-UAV-based real-time video surveillance system with stitching, in: Proc. 23rd ACM Int. Conf. Multimedia, 2015, pp. 261–270.

[117] Y. Wang, J. Farooq, Zero touch coordinated uav network formation for 360° views of a moving ground target in remote VR applications, in: Proc. IEEE Military Commun. Conf., 2022, pp. 950–955. doi:10.1109/MILCOM55135.2022.10017956.

[118] Z. Wang, J. Du, G. Li, X. Song, Control of panoramic video flow for UAV, in: Proc. Int. Conf. Social Comput. Digit. Economy, 2021, pp. 105–107.

[119] J. Qian, J. Wang, S. Jin, Configurable virtual MIMO via UAV swarm: Channel modeling and spatial correlation analysis, China Commun. 19 (9) (2022) 133–145.

[120] L. Teng, G. Zhai, Y. Wu, X. Min, W. Zhang, Z. Ding, C. Xiao, QoE driven VR 360° video massive MIMO transmission, IEEE Trans. Wireless Commun. 21 (1) (2022) 18–33. doi:10.1109/TWC.2021.3093305.

[121] Y. Ding, D. Jiang, J. Huang, L. Xiao, S. Liu, Y. Tang, H. Dai, QoE-aware power control for UAV-aided media transmission with

reinforcement learning, in: Proc. IEEE Global Commun. Conf., 2019, pp. 1–6. doi:10.1109/GLOBECOM38437.2019.9014064.

[122] R. Shirey, S. Rao, S. Sundaram, Optimizing quality of experience for long-range UAS video streaming, in: Proc. IEEE/ACM 29th Int. Symp. Quality Service, 2021, pp. 1–10. doi:10.1109/IWQOS52092.2021.9521330.

[123] J. Yang, J. Luo, D. Meng, J.-N. Hwang, QoE-driven resource allocation optimized for uplink delivery of delay-sensitive VR video over cellular network, IEEE Access 7 (2019) 60672–60683. doi:10.1109/ACCESS.2019.2915370.

[124] S. Colonnese, A. Carlesimo, L. Brigato, F. Cuomo, QoE-aware UAV flight path design for mobile video streaming in HetNet, in: Proc. IEEE 10th Sensor Array Multichannel Signal Process. Workshop, 2018, pp. 301–305.

[125] M. Tang, V. W. Wong, Online bitrate selection for viewport adaptive 360-degree video streaming, IEEE Trans. Mobile Comput. 21 (7) (2022) 2506–2517. doi:10.1109/TMC.2020.3038710.

[126] Z. Shu, T. Taleb, A novel qos framework for network slicing in 5g and beyond networks based on sdn and nfv, IEEE Network 34 (3) (2020) 256–263. doi:10.1109/MNET.001.1900423.

[127] L. A. b. Burhanuddin, X. Liu, Y. Deng, U. Challita, A. Zahemszky, QoE optimization for live video streaming in UAV-to-UAV communications via deep reinforcement learning, IEEE Trans. Veh. Technol. 71 (5) (2022) 5358–5370. doi:10.1109/TVT.2022.3152146.

[128] W. Huang, L. Ding, G. Zhai, X. Min, J.-N. Hwang, Y. Xu, W. Zhang, Utility-oriented resource allocation for 360-degree video transmission over heterogeneous networks, Digit. Signal Process. 84 (2019) 1–14.

[129] C. Perfecto, M. S. Elbamby, J. D. Ser, M. Bennis, Taming the latency in multi-user VR 360°: A QoE-aware deep learning-aided multicast framework, IEEE Trans. Commun. 68 (4) (2020) 2491–2508. doi:10.1109/TCOMM.2020.2965527.

[130] Y. Liu, C. Zhu, X. Deng, P. Guan, Z. Wan, J. Luo, E. Liu, H. Zhang, UAV-aided urban target tracking system based on edge computing, CoRR abs/1902.00837 (2019) 1–6.

[131] S. Park, S. R. Das, Cross-layer scheduling in QUIC and multipath QUIC for 360-degree video streaming, in: Proceedings of the IEEE Wireless Communication and Networking Conference, 2024, p. 1–6.

[132] A. Langley, A. Riddoch, A. Wilk, A. Vicente, C. Krasic, D. Zhang, F. Yang, F. Kouranov, I. Swett, J. Iyengar, J. Bailey, J. Dorfman, J. Roskind, J. Kulik, P. Westin, R. Tenneti, R. Shade, R. Hamilton, V. Vasiliev, W.-T. Chang, Z. Shi, The quic transport protocol: Design and internet-scale deployment, in: Proceedings of the Conference of the ACM Special Interest Group on Data Communication, SIGCOMM '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 183–196. doi:10.1145/3098822.3098842.
URL https://doi.org/10.1145/3098822.3098842

[133] E. Calvanese Strinati, S. Barbarossa, 6G networks: Beyond Shannon towards semantic and goal-oriented communications, Computer Networks 190 (2021) 107930.

[134] L. Qiao, M. B. Mashhadi, Z. Gao, C. H. Foh, P. Xiao, M. Bennis, Latency-aware generative semantic communications with pre-trained diffusion models (2024). arXiv:2403.17256.

[135] J. Devlin, M.-W. Chang, K. Lee, K. N. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018.
URL https://arxiv.org/abs/1810.04805

[136] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, M. Debbah, Large language models for telecom: The next big thing?, arXiv preprint arXiv:2306.10249 (2023).

[137] S. Vemprala, R. Bonatti, A. Bucker, A. Kapoor, ChatGPT for robotics: Design principles and model abilities, Tech. Rep. MSR-TR-2023-8, Microsoft (February 2023).
URL https://www.microsoft.com/en-us/research/publication/chatgpt-for-robotics-design-principles-and-model-abilities/

[138] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, Y. Cao, SimVLM: Simple visual language model pretraining with weak supervision, in: International Conference on Learning Representations, 2022.
URL https://openreview.net/forum?id=GUrhfTuf_3

[139] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, Flamingo: a visual language model for few-shot learning, in: A. H. Oh, A. Agarwal, D. Belgrave, K. Cho (Eds.), Advances in Neural Information Processing Systems, 2022.
URL https://openreview.net/forum?id=EbMuimAbPbs

[140] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023). arXiv:2301.12597.

[141] G. Delétang, et al., Language modeling is compression, arXiv preprint arXiv:2309.10668 (2023).

[142] N. Sehad, L. Bariah, W. Hamidouche, H. Hellaoui, R. Jäntti, M. Debbah, Generative ai for immersive communication: The next frontier in internet-of-senses through 6g, arXiv preprint arXiv:2404.01713 (2024).

[143] H. Zou, Q. Zhao, Y. Tian, L. Bariah, F. Bader, T. Lestable, M. Debbah, TelecomGPT: A framework to build telecom-specfic large language models, arXiv preprint arXiv:2407.09424v1 (2024).

Mohit K. Sharma (Member, IEEE) is a Senior Researcher at the Digital Sciences Research Center at the Technology Innovation Institute (TII), Abu Dhabi. He obtained the M.Tech and Ph. D. degrees from the Indian Institute of technology, Guwahati, India, and the Indian Institute of Science, Bangalore, India, respectively. From 2018 to 2019, he was a Postdoctoral Researcher at CentraleSupelec, Paris, France. From 2020 to 2022, he was a Research Scientist at the Institute for Infocomm Research at the Agency of Science, Technology, and Research (A*STAR), Singapore. Since, 2022 he is with the AI and Digital Sciences Research center at the TII, Abu Dhabi. His research interests are in developing physical layer techniques for the next generation wireless communication systems, and applications of Artificial Intelligence techniques to the design of wireless communication systems.

Ibrahim Farhat(brahim.farhat@tii.ae) is a Researcher at the Digital Sciences Research Center at the Technology Innovation Institute (TII), Abu Dhabi. Was born in Eljem, Tunisia, in 1993. He received the engineering degree in communication systems and computer science from SUP'COM school of engineering, Tunis, in 2018. In 2019, he joined the Institute of Electronic and Telecommunication of Rennes (IETR), Rennes, and became a member of the hardware team, where he received he's Ph.D. degree. Since, 2023 he is with the AI and Digital Sciences Research center at the TII, Abu Dhabi. His research interests are in developing novel AI-based techniques for implicit 3D scene representation and integrating to VR environments.

Chen-Feng Liu (Member, IEEE) received the B.S. degree from National Tsing Hua University, Hsinchu, Taiwan, in 2009, the M.S. degree in communications engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2011, and the Ph.D. degree in communications engineering from the University of Oulu, Oulu, Finland, in 2021. In 2012, he joined Academia Sinica, Taipei, Taiwan, as a Research Assistant. In 2022 and 2023, he was a Researcher with Technology Innovation Institute, Abu Dhabi, UAE. He is currently a Postdoctoral Research Associate with the New Jersey Institute of Technology, Newark, NJ, USA. His current research interests include 6G communications, ultra-reliable low-latency communications, and immersive video streaming. He has served as a member of the Technical Program Committee in a number of international conferences.

Nassim Sehad(nassim.sehad@aalto.fi) obtained the Bachelor of Science (B.Sc) diploma in the field of telecommunication in 2018 and the diploma of master in the field of networks and telecommunication in September 2020, from the University of Sciences and Technology Houari Boumediene (U.S.T.H.B), Algiers, Algeria. Since 2020 to September 2021 he joined a MOSA!C laboratory at Aalto University Finland as an assistant researcher. Since 2021 till now he joined the Department of Information and Communications Engineering (DICE), Aalto University, Finland, as a doctoral student. His main research topics of interest are multi-sensory multimedia, IoT, cloud computing, networks and AI.

Wassim Hamidouche(wassim.hamidouche@tii.ae) is a Principal Researcher at Technology Innovation Institute (TII) in Abu Dhabi, UAE. He also holds the position of Associate Professor at INSA Rennes and is a member of the Institute of Electronics and Telecommunications of Rennes (IETR), UMR CNRS 6164. He earned his Ph.D. degree in signal and image processing from the University of Poitiers, France, in 2010. From 2011 to 2012, he worked as a Research Engineer at the Canon Research Centre in Rennes, France. Additionally, he served as a researcher at the IRT b<>com research Institute in Rennes from 2017 to 2022. He has over 180 papers published in the field of image processing and computer vision. His research interests encompass various areas, including video coding, the design of software and hardware circuits and systems for video coding standards, image quality assessment, and multimedia security.

Mérouane Debbah(FELLOW IEEE) Mérouane Debbah is a researcher, educator and technology entrepreneur. Over his career, he has founded several public and industrial research centers, start-ups and is now Professor at Khalifa University of Science and Technology in Abu Dhabi and founding Director of the KU 6G Research Center. He is a frequent keynote speaker at international events in the field of telecommunication and AI. His research has been lying at the interface of fundamental mathematics, algorithms, statistics, information and communication sciences with a special focus on random matrix theory and learning algorithms. In the Communication field, he has been at the heart of the development of small cells (4G), Massive MIMO (5G) and Large Intelligent Surfaces (6G) technologies. In the AI field, he is known for his work on Large Language Models, distributed AI systems for networks and semantic communications. He received multiple prestigious distinctions, prizes and best paper awards (more than 40 IEEE best paper awards) for his contributions to both fields. He is an IEEE Fellow, a WWRF Fellow, a Eurasip Fellow, an AAIA Fellow, an Institut Louis Bachelier Fellow and a Membre émérite SEE.