Edge-Based Adaptive Services for Interactive and UAV-Based Systems in Softwarized 5G/6G Networks

Andrea Caruso
University of Catania
CNIT Research Unit
Catania, Italy
andrea.caruso@phd.unict.it

Christian Grasso
University of Catania
CNIT Research Unit
Catania, Italy
christian.grasso@unict.it

Giovanni Schembra
University of Catania
CNIT Research Unit
Catania, Italy
giovanni.schembra@unict.it

***Abstract*—Softwarized 5G/6G networks are fostering the emergence of adaptive services operating over highly dynamic and heterogeneous environments. In this context, edge computing enables the distribution of processing and control, allowing applications to react to network variability, resource availability, and system dynamics in real time.**

This PhD research investigates network-aware edge-based adaptive mechanisms driven by AI to support two representative classes of services. The first focuses on immersive interactive applications, such as 360° video streaming and cloud gaming, where QoE depends on the joint management of video delivery, user interaction, and network conditions. The second addresses distributed AI-driven systems deployed over UAV networks, where sensing, communication, and inference must be coordinated under mobility and resource constraints.

The proposed approach adopts a unified perspective in which edge and far-edge resources, AI-driven adaptation, and Digital Twin models are combined into a common control methodology for efficient and responsive service behavior. The effectiveness of this methodology is validated through both simulation and real-world experimental setups, showing its ability to cope with dynamic conditions while maintaining performance and resource efficiency.

***Index Terms*—Network Softwarization, Edge Computing, AI-driven Adaptation, Digital Twin Networks, QoE-aware Services**

I. Introduction

Softwarized 5G/6G networks are enabling a new class of adaptive services operating over heterogeneous and dynamic environments, where application performance depends on the tight interaction between data generation, processing, delivery, and network resource availability. In this context, edge computing plays a central role by bringing computation closer to data sources and end users, enabling low-latency responsiveness, scalable resource utilization, and context-aware service adaptation. At the same time, network softwarization, slicing, and emerging network-exposure mechanisms make network conditions and capabilities explicit inputs for application-level control [1], [2]. Among these services, immersive interactive applications, such as 360° video streaming and cloud gaming, pose stringent requirements due to the strong coupling between user behavior, system response, and network conditions. In such scenarios, Quality of Experience (QoE) is influenced by multiple factors, including viewport dynamics, end-to-end latency, and bandwidth variability. These characteristics make traditional delivery mechanisms inadequate, motivating the adoption of adaptive encoding strategies, viewport-aware processing, and edge-assisted control mechanisms capable of reacting to rapid system variations [3]–[7]. The same need for network-aware adaptation also emerges in distributed systems based on UAVs and Flying Ad-hoc Networks (FANETs), which introduce additional challenges related to highly dynamic topologies, intermittent connectivity, and constrained onboard resources [8]–[11]. In these environments, data acquisition, processing, and decision-making must be distributed across heterogeneous nodes, requiring efficient coordination between sensing, communication, and computation. Recent research has explored the integration of edge and far-edge computing, AI-driven adaptation, and Network Digital Twins (NDTs) to support real-time data processing, system monitoring, and predictive optimization [12]–[16]. Despite these advances, existing solutions typically address interactive services and UAV-based systems independently, often focusing on specific layers (e.g., encoding, networking, or inference) without providing a unified view of how adaptation mechanisms can be designed across heterogeneous domains. Moreover, network information is frequently treated as an external constraint rather than as an exposed and actionable component of the control loop. In particular, the joint optimization of communication, computation, and system dynamics at the edge remains an open challenge.

This PhD work addresses this gap by investigating AI-driven edge-based adaptation in softwarized networks through two complementary directions: QoE-aware interactive services, leveraging adaptive encoding and edge-assisted processing [17]–[21], and distributed AI-driven UAV systems, exploiting edge and far-edge resources for flexible data processing, continuous model adaptation, and system-level coordination [22]–[25]. Although targeting different domains, both directions rely on a common network-aware control loop that uses application, context, and network information to drive AI-based decisions across edge and far-edge resources, thus providing a unified perspective on adaptive service design.

ISBN 978-3-903176-82-9

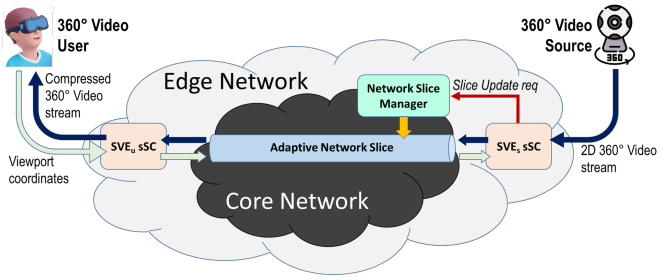

Figure 1: Proposed system for interactive services

II. Research Questions

This work is driven by two complementary research questions:

RQ1: How to design QoE-aware closed-loop adaptation for interactive services such as VR streaming and cloud gaming under dynamic network conditions?

RQ2: How to enable efficient edge-based adaptation of distributed AI-driven pipelines in highly dynamic and resource-constrained systems, such as UAV networks?

III. Methodology and Current Status

The proposed approach adopts a network-aware edge-centric design, where AI-driven adaptation captures the interaction between application requirements, network dynamics, and distributed computing resources. This methodology is applied to interactive services and distributed UAV-based systems, whose reference architectures are illustrated in Fig. 1 and Fig. 2, respectively. Both scenarios distribute intelligence across edge and far-edge resources through closed-loop mechanisms that use application- and network-level observations to update service decisions at runtime.

A. Interactive Services

The first research direction targets 360° video streaming and VR cloud gaming over edge-enabled networks. The architecture is based on a split Service Chain, where the Smart Video Encoder (SVE) is decomposed into source-side (SVE_s) and user-side (SVE_u) edge sub-chains (Fig. 1), enabling video processing and control close to both the content source and the user. At the source side, 360° frames are projected into a 2D equirectangular format and partitioned into concentric tile zones, enabling viewport-aware compression with higher quality near the viewport and stronger compression in peripheral regions. A DRL agent dynamically selects tile-level parameters based on bandwidth estimates and viewport updates.

The encoding pipeline is implemented within the SVE, where the DRL agent selects compression configurations under bandwidth constraints. The system also interfaces with a Network Slice Manager, which monitors slice-level conditions and triggers reconfiguration upon bandwidth or latency Service Level Agreement (SLA) violations. At the user side, viewport information is collected through a custom Android application running on a smartphone embedded in a VR visor, logging azimuth, pitch, and roll, and transmitted via MQTT to the source. The user-side module also supports bandwidth estimation and video playback, thus closing the control loop between user behavior, network state, and source-side encoding decisions. Due to bidirectional latency, the viewport used during encoding may differ from the actual one at playback. This effect is captured through a perceived-quality model that weights each tile according to its compression level and spatial relevance to the user's field of view, guiding the DRL agent toward delay-robust configurations. The framework is extended to cloud gaming by integrating input synchronization: commands are transmitted via MQTT, video via RTSP, and delay estimation, buffering, and frame-command alignment are used to compensate controlled latency. Three AI agents with different accuracy levels support reproducible evaluation. Experimental results show that DRL-based encoding outperforms heuristic and alternative learning-based approaches, increasing perceived quality from about 40 dB at 75 Mbps to over 48 dB at 150 Mbps, with gains up to 16% over differential compression. Performance remains stable under delays up to 1 s, with controlled frame sizes and near-zero bandwidth violations. In cloud gaming, input-delay variability degrades performance across all skill levels, especially for more skilled agents, highlighting the impact of residual jitter.

B. Distributed AI-Driven UAV Systems

The second research direction focuses on distributed AI-driven UAV systems for monitoring and safety applications. The architecture, depicted in Fig. 2, follows a multi-layer design integrating onboard sensing, aerial far-edge processing, ground-edge intelligence, and a Digital Twin platform. At the UAV level, heterogeneous sensing and computing capabilities are deployed, including RGB, multispectral, LiDAR, and thermal data acquisition, supported by embedded onboard units interfaced through UAV SDKs. The software stack extracts telemetry and manages video acquisition. Telemetry and video streams are transmitted through distinct channels: MQTT for lightweight telemetry exchange and RTSP for continuous video streaming, enabling differentiated handling of flows with diverse latency, reliability, and bandwidth requirements. At the aerial far-edge layer, incoming video streams are processed through a data distillation pipeline that filters redundant frames using similarity metrics, reducing bandwidth consumption while preserving scene dynamics. Additional modules regulate frame rate according to available bandwidth and may execute lightweight inference for low-latency responses. Filtered data are then delivered to the ground edge, where dataset creation and model lifecycle management are performed. In particular, a YOLOE-based annotation pipeline generates labeled datasets from streaming data through text-driven prompts, supporting continuous enrichment with limited manual effort. The generated datasets feed a model training and optimization pipeline. A dual deployment strategy is adopted: high-capacity models

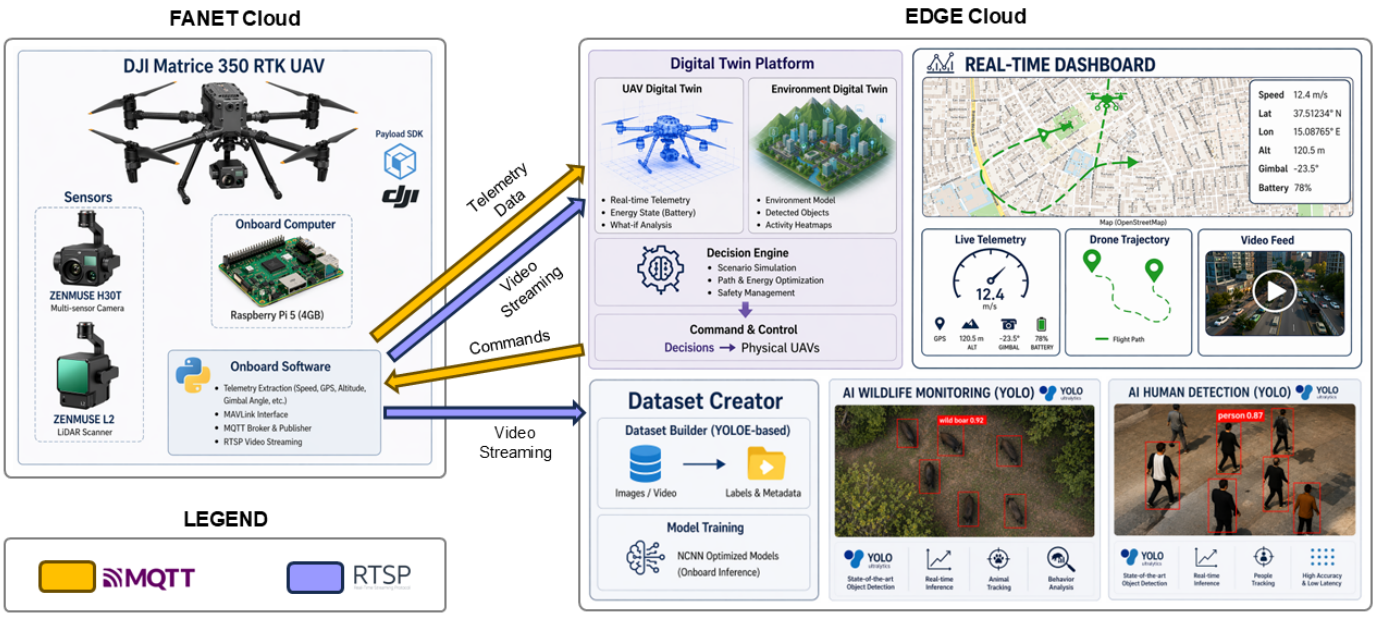

Figure 2: Architecture of the UAV-based distributed AI system with Digital Twin integration and edge-cloud orchestration

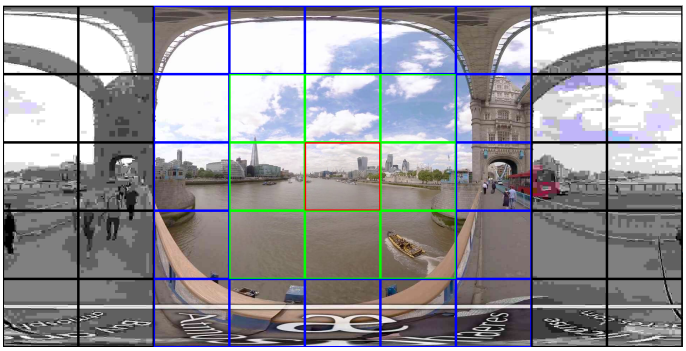

Figure 3: Tile-based encoding with differentiated compression

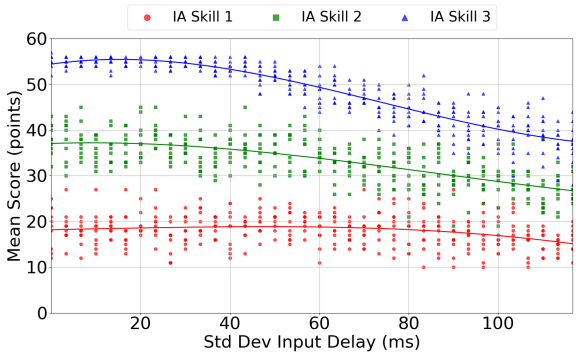

Figure 4: Impact of the Standard Deviation of Input Delay on the three AI players after mean input delay compensation

are executed at the ground edge for accurate inference over complete data streams, while optimized lightweight models are deployed onboard UAVs to enable real-time inference under strict computational and energy constraints. Model optimization includes conversion into efficient formats suitable for embedded execution, enabling low-latency processing directly on the UAV. The Digital Twin platform integrates telemetry and perception outputs into synchronized representations of both the UAV state and the surrounding environment. The UAV Digital Twin models dynamic parameters such as position, motion, and operational status, while the Environment Digital Twin aggregates detected objects, their spatial distribution, and temporal evolution. These representations are continuously updated and support real-time visualization, monitoring, and decision-making through a geospatial interface. The overall system operates as a closed loop, where sensed data are filtered at the far-edge, transformed into datasets and models at the ground edge, and reintegrated into the operational workflow through adaptive model deployment and Digital Twin updates. This enables continuous interaction between sensing, learning, decision-making, and network-aware resource usage. Experimental results show that similarity-based filtering significantly reduces transmitted data while preserving scene dynamics, and that the automated annotation pipeline produces high-quality datasets with negligible errors. Lightweight models trained on these data achieve precision and recall of 0.955, with mAP of 0.961 (IoU 0.5) and 0.832 (IoU 0.5–0.95). The hybrid inference strategy balances latency and accuracy, while RGB–multispectral fusion improves classification in complex scenarios.

IV. Conclusions and Future Work

The presented results demonstrate that network-aware edge-based adaptive mechanisms can effectively support interactive and distributed AI-driven services under dynamic conditions, enabling service behavior to be adapted across edge and far-edge resources according to application, context, and network information. For interactive services, the results show that DRL-based adaptive encoding effectively improves perceived quality while maintaining strict bandwidth compliance. The combination of viewport-aware compression and delay-aware modeling ensures robustness to network variability, while the cloud gaming extension highlights that, beyond average latency, delay variability remains a key factor affecting interaction consistency. For UAV-based systems, the integration of data filtering, automated dataset generation, and hybrid inference enables efficient and scalable AI pipelines. The YOLOE-based annotation process supports continuous model adaptation with limited human intervention, while the distributed execution across onboard and edge resources allows balancing latency, accuracy, and communication overhead. The adoption of Digital Twins further enhances system observability and supports real-time decision-making. Future work will extend this direction along multiple axes. In the interactive domain, ongoing efforts move from viewport-driven to content-aware encoding, exploiting saliency-aware tile compression and spatio-temporal attention prediction at the edge without requiring client feedback. Further work will address multi-user and multi-UAV scenarios, standardized network-exposure interfaces, refined edge–cloud orchestration, and predictive Digital Twin models for proactive adaptation.

Acknowledgments

The research was partially supported by the research project "Programma ricerca di Ateneo UNICT 2024-26 NOVA - Network Optimization and Vulnerability Assessment" of the University of Catania.

References

[1] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6g: A comprehensive survey," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 334–366, 2021.

[2] S. Kianpisheh and T. Taleb, "A survey on in-network computing: Programmable data plane and technology specific applications," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 701–761, 2022.

[3] X. Corbillon, G. Simon, A. Devlic, and J. Chakareski, "Viewport-adaptive navigable 360-degree video delivery," in *2017 IEEE international conference on communications (ICC)*, pp. 1–7, IEEE, 2017.

[4] L. Yu, T. Tillo, and J. Xiao, "Qoe-driven dynamic adaptive video streaming strategy with future information," *IEEE Transactions on Broadcasting*, vol. 63, no. 3, pp. 523–534, 2017.

[5] T. Stockhammer, "Dynamic adaptive streaming over http– standards and design principles," in *Proceedings of the second annual ACM conference on Multimedia systems*, pp. 133–144, 2011.

[6] S. Petrangeli, G. Simon, H. Wang, and V. Swaminathan, "Dynamic adaptive streaming for augmented reality applications," in *2019 IEEE International Symposium on Multimedia (ISM)*, pp. 56–567, IEEE, 2019.

[7] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE communications surveys & tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[8] C. Grasso, R. Raftopoulos, and G. Schembra, "Slicing a fanet for heterogeneous delay-constrained applications," *Comput. Commun.*, vol. 195, p. 362–375, Nov. 2022.

[9] C. Grasso, R. Raftopoulos, and G. Schembra, "Multi-agent deep reinforcement learning in flying ad-hoc networks for delay-constrained applications," *Procedia Computer Science*, vol. 203, pp. 69–78, 2022.

[10] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in uav communication networks," *IEEE communications surveys & tutorials*, vol. 18, no. 2, pp. 1123–1152, 2015.

[11] S. Hayat, E. Yanmaz, and R. Muzaffar, "Survey on unmanned aerial vehicle networks for civil applications: A communications viewpoint," *IEEE communications surveys & tutorials*, vol. 18, no. 4, pp. 2624–2661, 2016.

[12] J. Li, Q. He, X. Wang, A. Hawbani, K. Yu, Y. Bi, and L. Zhao, "Uav-assisted microservice mobile edge computing architecture: Addressing post-disaster emergency medical rescue," *IEEE Transactions on Computers*, 2025.

[13] Y. Zhang, Y. Gong, and Y. Guo, "Energy-efficient resource management for multi-uav-enabled mobile edge computing," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 8, pp. 12026–12037, 2024.

[14] S. Wu, N. Chen, A. Xiao, P. Zhang, C. Jiang, and W. Zhang, "Ai-empowered virtual network embedding: a comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 27, no. 2, pp. 1395–1426, 2024.

[15] D.-H. Tran, N. Waheed, Y. M. Saputra, X. Lin, C. T. Nguyen, T. S. Abdu, V. N. Vo, V.-Q. Pham, M. Alsenwi, A. B. M. Adam, *et al.*, "Network digital twin for 6g and beyond: An end-to-end view across multi-domain network ecosystems," *IEEE Open Journal of the Communications Society*, 2025.

[16] Y. Wu, K. Zhang, and Y. Zhang, "Digital twin networks: A survey," *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 13789–13804, 2021.

[17] A. Caruso, C. Grasso, R. Raftopoulos, and G. Schembra, "An adaptive closed-loop encoding vnf for virtual reality applications," in *2024 IEEE 10th International Conference on Network Softwarization (NetSoft)*, pp. 222–230, IEEE, 2024.

[18] A. Caruso and G. Schembra, "A vr 360°-video encoding framework with differentiated tile compression based on digital-twin technology," in *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 515–521, IEEE, 2024.

[19] A. Caruso and G. Schembra, "Impact of user's movements in an immersive 360° video streaming with differentiated compression," in *2024 IEEE 67th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 712–718, IEEE, 2024.

[20] V. Charpentier, G. Landi, E. Giannopoulou, J. Brenes, R. Frizzell, M. Iordache, C. Patachia, P. Demestichas, G. Baldoni, A. Caruso, *et al.*, "Utilizing the vital-5g platform to advance 5g standalone integration with vertical industries," *EURASIP Journal on Wireless Communications and Networking*, vol. 2025, no. 1, p. 27, 2025.

[21] J. Brenes, A. Caruso, P. G. Giardina, C. Grasso, G. Landi, L. Lossi, G. Schembra, and G. Scivoletto, "Experimenting over a federated 6g network infrastructure: Adaptive 360 video streaming," in *2024 IEEE 29th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pp. 1–6, IEEE, 2024.

[22] G. Davoli, C. Grasso, A. Caruso, W. Cerroni, G. Colajanni, L. Galluccio, and G. Schembra, "A marketplace approach for service-chain deployment in a multi-layer fanet edge-computing architecture," in *2025 Integrated Communications, Navigation and Surveillance Conference (ICNS)*, pp. 1–10, IEEE, 2025.

[23] A. Caruso, L. Galluccio, C. Grasso, M. Ignaccolo, G. Inturri, P. Leonardi, G. Schembra, and V. Torrisi, "Advancing urban traffic monitoring in smart cities: A field experiment with uav-based system for transport planning and intelligent traffic management," in *2025 Integrated Communications, Navigation and Surveillance Conference (ICNS)*, pp. 1–9, IEEE, 2025.

[24] A. Caruso, C. Grasso, R. Raftopoulos, and G. Schembra, "Falcon: Fanet-aware learning and digital twin control framework," *Computer Communications*, vol. 251, p. 108481, 2026.

[25] A. Caruso, "A uav-based hybrid human-ai training for wild-animal detection," in *HHAI 2025: Proceedings of the 4th International Conference on Hybrid Human-Artificial Intelligence*, vol. 408 of *Frontiers in Artificial Intelligence and Applications*, pp. 544–551, Amsterdam, The Netherlands: IOS Press, Sept. 2025.