Toward 6G-Enabled Mobile Vision Analytics for Immersive Extended Reality

Miao Zhang, Linfeng Shen, Xiaoqiang Ma, *Member IEEE*, Jiangchuan Liu, *Fellow IEEE*

***Abstract*—The fifth generation (5G) communication systems have seen initial success in boosting a broad spectrum of mobile networked applications. However, emerging applications, notably immersive eXtended Reality (XR), have already posed significant new challenges to today's 5G given their ultra-high expectations on data rate and latency. They also demand a deep integration of communication and computation for data analytics. In particular, by analyzing the vision data captured by mobile devices, Mobile Vision Analytics (MVA) facilitates understanding of the ambient environment in XR, a key to truly immersive and interactive experiences. In this article, we advocate MVA as a core service of the forthcoming 6G. We closely examine the potentials and challenges of 6G-enabled MVA and accordingly present an integrated framework with massively distributed 6G edge computing nodes to power ubiquitous vision analytics. We identify the critical design issues towards its implementation and shed light on a series of advanced enabling technologies.**

***Index Terms*—Mobile vision analytics, eXtended Reality (XR), 6G, Edge computing**

I. Introduction

Cyberspace has long had an ambitious goal — connecting the world, understanding the world, and interacting with the world, both physically and virtually, for human beings and machines, anytime and anywhere. This remained a dream a decade ago. With the unprecedented development in the Information and Communication Technology (ICT) sector in the past decade, however, it is now solid and reachable to a great extent, if not all.

The main driving forces behind the scene are advanced communication technologies (from 4G to 5G and beyond) and intelligent computing technologies, particularly deep neural networks (DNNs). The former interconnects ubiquitous devices for physical data acquisition and aggregation, and the latter offers a deep understanding of the physical world through large-scale neural networks of cascaded layers. They together have fostered a number of emerging applications, notably eXtended Reality (XR), a universal term inclusive of virtual reality (VR), augmented reality (AR), and mixed reality (MR). XR shows great application potential in sectors such as education, healthcare, and entertainment by seamlessly integrating the physical and digital worlds and allowing users to immerse themselves and interact with objects in the hybrid world. Among the many types of data, vision data captured by cameras of XR devices are undoubtedly the most informative for sensing and interaction. Vision analytics, which analyzes such data with vision DNNs to extract knowledge and insights for understanding the physical world, therefore, plays a crucial role in modern XR systems [1].

5G New Radio (NR) began its deployment in 2019, though its initial design dates back to earlier than 2010.[1] The International Telecommunication Union Radiocommunication Sector (ITU-R) has defined three main service types for 5G's advanced capabilities: Enhanced Mobile Broadband (eMBB), Ultra Reliable Low Latency Communications (URLLC), and Massive Machine Type Communications (mMTC). As a direct extension of 4G's broadband service, the data rate is maximized with eMBB, and XR is one of its main targeted applications. Unfortunately, reliably analyzing camera streams for latency-critical XR applications is non-trivial, as it requires running computationally intensive DNNs at high speeds. This is particularly challenging for *Mobile Vision Analytics* (MVA) (Fig. 1[2] and Fig. 2), where the videos are captured by resource-constrained mobile and wearable devices that cannot support high-accuracy in-situ analytics [1], [2]. 5G addresses this issue by introducing *Mobile Edge Computing* (MEC), allowing data to be offloaded to nearby edge servers for analysis.

MEC ignites the integration of communication, storage, and computation at the network edge. Nevertheless, immersive XR with MVA can hardly be fulfilled by today's eMBB, which is mostly optimized for downlink (DL), not uplink (UL). Even with MEC servers capable of running state-of-the-art DNNs in real time, shipping regular videos over ULs can be the latency bottleneck of the entire analytics pipeline [3], let alone much larger 360° and volumetric videos. URLLC and mMTC remain unavailable in most 5G deployments, and in fact, they will not co-exist when serving a user. Lacking of high-quality ULs, various workarounds have been proposed to reduce the amount of offloaded data, such as degrading video quality [4] and recovering with server-side super-resolution (SR) models, dynamic region of interest (RoI) encoding [5], and DNN model splitting [6]. Unfortunately, these efforts are often suboptimal as they need to sacrifice accuracy to achieve the desired latency. Without an offloading service that simultaneously offers ultra-broad bandwidth and ultra-low latency, particularly for the ULs, continuous and complete awareness of the physical world is hard to develop. We expect the forthcoming 6G to break the current barrier and realize ubiquitous MVA through its Tbps data rate and sub-millisecond latencies [7], [8].

Miao Zhang, Linfeng Shen, and Jiangchuan Liu (corresponding author) are with the School of Computing Science, Simon Fraser University, Burnaby, Canada.

Xiaoqiang Ma is with the CSIS Department, Douglas College, Canada.

[1]3GPP Release 15. https://www.3gpp.org/specifications-technologies [Accessed Mar 12, 2023]

[2]Cityscapes: https://www.cityscapes-dataset.com/; Microsoft COCO: https://cocodataset.org; DIODE: https://diode-dataset.org/ [Accessed Mar 12, 2023]

(a) Semantic Segmentation (data source: Cityscapes)

(b) Object detection (data source: Microsoft COCO)

(c) Human keypoint detection (data source: Microsoft COCO)

(d) Depth estimation (data source: DIODE)

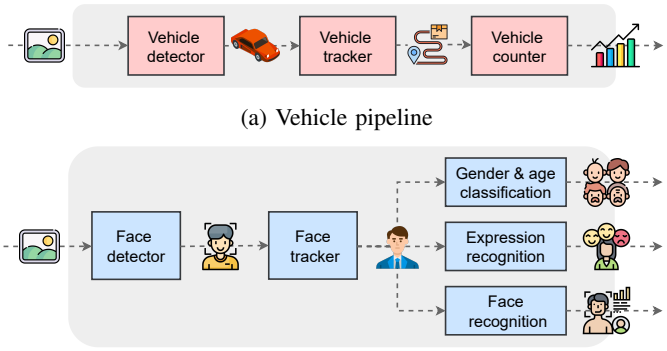
Fig. 1: Examples of basic MVA tasks.

We also advocate MVA as a core service in the 6G development, given its importance in bridging the virtual and physical worlds. We believe that ubiquitously deployed *Edge Computing Nodes* (ECNs) co-located with 6G base stations (BSes) will host various vision models to serve heterogeneous analytics requests from any mobile devices, regardless of their computing power (weak or strong), location (urban or rural), and mobility (static or high-speed). Collaborative MVA at scale will become true with the closer cooperation between ECNs. Despite promising, seamless integration with 6G involves a series of issues, including but not limited to image and video compression optimization, efficient resource management across ECNs, and user data security and privacy preservation.

In this article, we first review state-of-the-art works of MVA and identify its key challenges with today's communication and computing infrastructure. We then present an integrated design for MVA over 6G and discuss its critical components. We further demonstrate a range of advanced technologies that will facilitate the design for pervasive XR experiences across devices and geo-locations.

II. MVA and 6G: A Tale of Two Technologies

Though originated from different technological streams, immersive XR and mobile broadband have partnered up from the 3G era, with the former being the driving application and the latter being the underlying vehicle. Since then, they have evolved and fostered each other's growth. Together they will make profound changes to our lives in the foreseeable future. For example, conferences can be hosted in a hybrid physical and virtual world to hurdle the operational barriers caused by physical distance; engineers can monitor the work in progress and solve mechanical problems without physically visiting remote or dangerous sites. These all demand a deep understanding of the physical environments to reconstruct, synchronize, and extend real-world experiences in the virtual world. MVA, which allows machines to understand the physical world through cameras, serves as one driving force and fundamental building block for immersive XR. For instance,



(a) Vehicle pipeline

(b) Face pipeline

Fig. 2: Two representative advanced MVA pipelines.

depth estimation can help appropriately place virtual objects in the physical world, and hand tracking enables humans in the physical world to interact with virtual objects. We now present a reality check of the context, focusing on MVA for XR (as summarized in Table I) and the unique challenges of its implementation over today's mobile networks.

A. MVA for XR Perception

The resource demand for MVA is enormous (see two representative MVA pipelines in Fig. 2). For instance, the latest YOLOv7-E6E[3] model needs a powerful GPU, if not the most advanced (e.g., NVIDIA V100 Tensor Core, 14 FP32 TFLOPS), to achieve near real-time inference (36 FPS). Running it in real time is clearly beyond the processing capacity of today's best smartphone chips (e.g., A16 Bionic at 1.8 FP32 TFLOPS and Snapdragon 8 Gen 2 at 3.7 FP32 TFLOPS), not to mention other weaker mobile or wearable devices (e.g., AR/MR headsets and smart cameras). The power consumption associated with such models also leads to excessive heat dissipation for the devices. Model pruning and compression techniques have been suggested for the subpar devices [4];

[3]https://github.com/WongKinYiu/yolov7 [Accessed Mar 12, 2023]

TABLE I: Representative MVA solutions.

Existing work	Video type	Shipped data	Analytics tasks	Edge server
DeepDecision [4]	normal	compressed camera frames	object detection	single server
Liu et al. [5]	normal	compressed frame slices	object detection, human keypoint detection	single server
Meng et al. [9]	normal	compressed camera frames	object detection	single server
ANS [6]	normal	intermediate DNN inference results	object detection	single server
Elf [2]	high-resolution	compressed frame partitions	instance segmentation, object classification, pose estimation	multiple servers
DeepMix [10]	3D videos	compressed RGB frames	3D object detection	single server

the reduced accuracy, unfortunately, can significantly hurt the quality of experience (QoE) of XR users.

To overcome the resource constraints, it is necessary to offload certain analytics workloads from weak front-end mobile or wearable devices to powerful back-end edge/cloud servers. DeepDecision [4] is an earlier attempt that employs a measurement-driven mathematical framework to determine an optimal offloading strategy for AR applications by considering model accuracy, video quality, battery constraints, and network conditions. To reduce the network transmission latency in edge-assisted mobile AR applications, Liu *et al.* [5] suggest dynamic RoI encoding to adjust the offloading data rate and power consumption, which works together with on-device object tracking to maintain a high detection accuracy. Meng *et al.* [9] revisit the canonical design of edge-assisted AR and find that video compression plus on-device object tracking show satisfactory effectiveness in edge-assisted object detection. Another line of research considers DNN model splitting instead of using standalone models on mobile devices or edge servers for inference. For example, with the observation that the intermediate data size of DNN is much smaller than the original input data, ANS [6] automatically partitions a DNN into two parts, with the front-end part running on the mobile device and the back-end part running on the edge server.

As camera technology evolves and hardware costs continue to decrease, ultra high definition (UHD, e.g., 4K and 8K) videos at high frame rates (100+ FPS) will be readily captured by the latest mobile devices. This paves the way to realistic scene reconstruction and fine-grained object identification in XR, which ideally expects a 32K+ resolution.[4] 360° videos further cover an omnidirectional *field of view* (FoV) beyond the limited FoV of traditional videos. Analyzing such videos can provide users with full situational awareness without blind spots. Nevertheless, under the same perceived quality, 360° videos can be 4× to 6× larger than regular videos, making MVA increasingly data-intensive and resource-intensive. Its unique spherical geometry also challenges vision models initially designed for and trained on flat images [11].

Newer data capture devices, such as Light Detection and Ranging (LiDAR) and hyperspectral cameras, have further enhanced our perception of the environment with rich depth or spectral information, laying the foundation for truly immersive experiences and interactions [10]. For instance, 3D mobile vision can offer higher accuracy than its 2D counterpart by detecting distant small objects and mitigating the occlusion issues [1]. However, today's communication and computing infrastructures are not ready for real-time 3D or hyperspectral vision analytics due to the dramatically increased resource demands for processing additional depth or spectral data.

There have been initial attempts toward MVA with these new-generation visual data that contain much richer information. To accommodate the huge data volume of UHD, Elf [2] leverages multiple edge servers with data-parallel content-aware frame partitioning for offloading acceleration. For 3D MVA, DeepMix [10] presents a workaround that offloads only RGB images to the edge for 2D object detection and then fine-tunes the results on the mobile device with depth information to boost accuracy.

B. MVA over 5G: Reality Check

MEC is regarded as a key technology and architectural concept to enable the evolution to 5G [12]. Together with network slicing, it is expected to accommodate a wide spectrum of computation-intensive applications that could not be executed on mobile devices, including MVA. Assuming the refresh rate of a camera is 60 Hz, the underlying vision analytics should be completed within 16.7 ms to enable real-time interaction. Fig. 3 shows a canonical MEC-based MVA implementation where the mobile device sends video frames to the edge server for vision analysis. The end-to-end (E2E) latency for analyzing a frame consists of four parts, the upload time (UT) of sending the frame to the edge server, the server-side inference time (IT), the analysis results download time (DT), and the post-processing and rendering time (PRT) on the mobile device. For most MVA tasks (e.g., object detection and hand tracking), DT can be neglected because the analysis results are only several bytes [12] while the DL bandwidth is relatively ample. IT and PRT are highly hardware-dependent and determined by the computing power of edge servers and mobile devices, respectively. Therefore, network-related latency optimization for MVA should focus on UT.

Unfortunately, the user-perceived UL speed of 5G is unsatisfactory. According to a recent report,[5] although the average 5G DL speed of T-Mobile, a major US operator, achieves 171 Mbps, the average 5G upload speed only reaches 17.8 Mbps. At this speed, it would take several seconds to upload even a highly-compressed 4K frame, which is significantly longer than the tolerant latency of MVA (e.g., 16.7 ms). Furthermore,

[4]http://www.clarkvision.com/articles/human-eye/ [Accessed Mar 12, 2023]

[5]https://www.opensignal.com/reports/2022/07/usa/mobile-network-experience-5g [Accessed Mar 12, 2023]

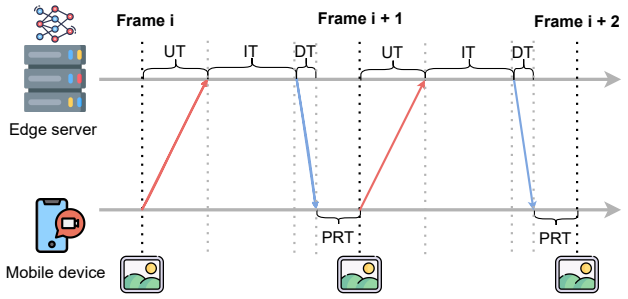


Fig. 3: The offloading latency breakdown for a canonical edge-assisted MVA architecture.

as XR becomes pervasive, we have to simultaneously track numerous objects from massive mobile devices, together with many other tasks in XR scene construction, integration, and interactions. MVA at scale indeed demands eMBB, URLLC, and even mMTC, which are not simultaneously supported by 5G. To deliver on the long-held immersive and interactive promise, new-generation mobile broadband beyond 5G should be put forward on the agenda.

C. MVA as a 6G Core Service: Design Space

As of now, there is no universally-accepted standard for 6G from the government or industry. Nonetheless, without a doubt, 6G will be application-driven and human-centric, supporting applications beyond current mobile use cases [8], [13]. Advanced VR and AR, pervasive intelligence, and Internet of everything (IoE) have been frequently cited as 6G's targets, all of which are essential components in immersive XR. With Terahertz and millimeter wave (mmWave) communications, 6G is anticipated to provide Tbps data rate, sub-millisecond latency, and ubiquitous connections with high efficiency [7], [8], [13]. Beyond these exciting measures from the communication perspective, we envision the following functions of 6G for intelligent and interactive applications, in particular, MVA as a key enabler towards immersive XR:

Inference task offloading as a network primitive: Mobile network operators (MNOs) are anticipated to adopt flexible, decentralized business models for 6G with spectrum sharing, infrastructure sharing, and intelligent automated management. In this context, we advocate inference task offloading as a network primitive in 6G.

Unlike previous generations of mobile networks that largely focus on optimizing the DL-centric QoE, 6G is expected to provide high-throughput (Tbps) and low-latency (0.1 ms) wireless UL as well. As such, in 6G, heavy, bursty video data from multitudinous mobile devices can be uploaded to the densely deployed high-performance edge servers through high-speed ULs and analyzed by application-customized and high-accuracy vision models in milliseconds or below. Also, note that 5G MEC does not differentiate inference tasks from other data traffic nor consider the pipelined operation of inference tasks. Since 6G will be application-driven and empower massively distributed intelligence, these should be factored into the 6G MEC design. The design should also be open enough to seamlessly accommodate the latest analytics tools and edge computing platforms, e.g., EdgeX Foundry[6].

Mobility-aware and mobility-oblivious computing: Recent measurement results of 5G have revealed that the network throughput exhibits substantial variations while walking or driving [3], [14]. The UL throughput can often drop to < 10 Mbps while driving [3], and the prolonged E2E latency in MVA can severely degrade the XR QoE. In the indoor environment, 5G mmWave may incur high path loss, as obstacles like walls easily block high-frequency signals.

6G-enabled MVA should be mobility-aware in network-wide resource management and cooperation, providing stable and satisfactory analytics service for indoor and outdoor users. Additionally, 6G is expected to encompass various mobile access technologies, possibly including the global-range low Earth orbit (LEO) satellites and the short-range backscatter. Its cell coverage will also be more versatile, including densely deployed tiny cells (covering tens of meters [13]). Apparently, handoff across cells and communication technologies will be more frequent. Thus, it is necessary to mask the underlying network heterogeneity and dynamics so that users in the same virtual world but with diverse network access conditions can have a uniform experience. Note that the movement in the physical world may be inconsistent with that in the virtual world, e.g., a presenter in a virtual conference room may be physically on a high-speed train.

Pervasive collaboration and integration: 5G-empowered MVA has primarily focused on optimizing QoE for a single camera stream. They tend to understand and track sporadic real-world objects appearing in a limited FoV for a short time. With the recent advances in video capture and analytics, the next wave of XR will embrace large-scale collaborative analytics of multiple camera streams or even mass camera arrays. Multitudinous objects or things of the physical world will be continually identified and tracked over extensive temporal and spatial ranges. The outcomes can then be integrated to reconstruct a holistic digital world, closing the gap between the physical and virtual worlds for truly immersive XR. This paradigm shift calls for tight collaborations across geo-distributed edge nodes, as each node may only be able to track an object within a certain space-time window. In other words, together with the network primitive for offloading and mobility awareness, 6G MEC should offer aggregated and transparent resources across its edges to satisfy the mobile users' continuing MVA requests.

III. 6G-Enabled MVA: An Integrated Framework

Given our ultimate goal of understanding and intellectualizing the physical world and further bridging the gap between the physical and virtual worlds, forward-thinking to 6G and its driving applications is necessary from now on. We believe that MVA will be the key to realizing our expectations of XR and envision a 6G-enabled "MVA as a service" framework, as demonstrated in Fig. 4. This framework allows users to enjoy seamless MVA service from an integrated terrestrial, airborne, and satellite access network.

[6]https://www.edgexfoundry.org/ [Accessed Mar 12, 2023]

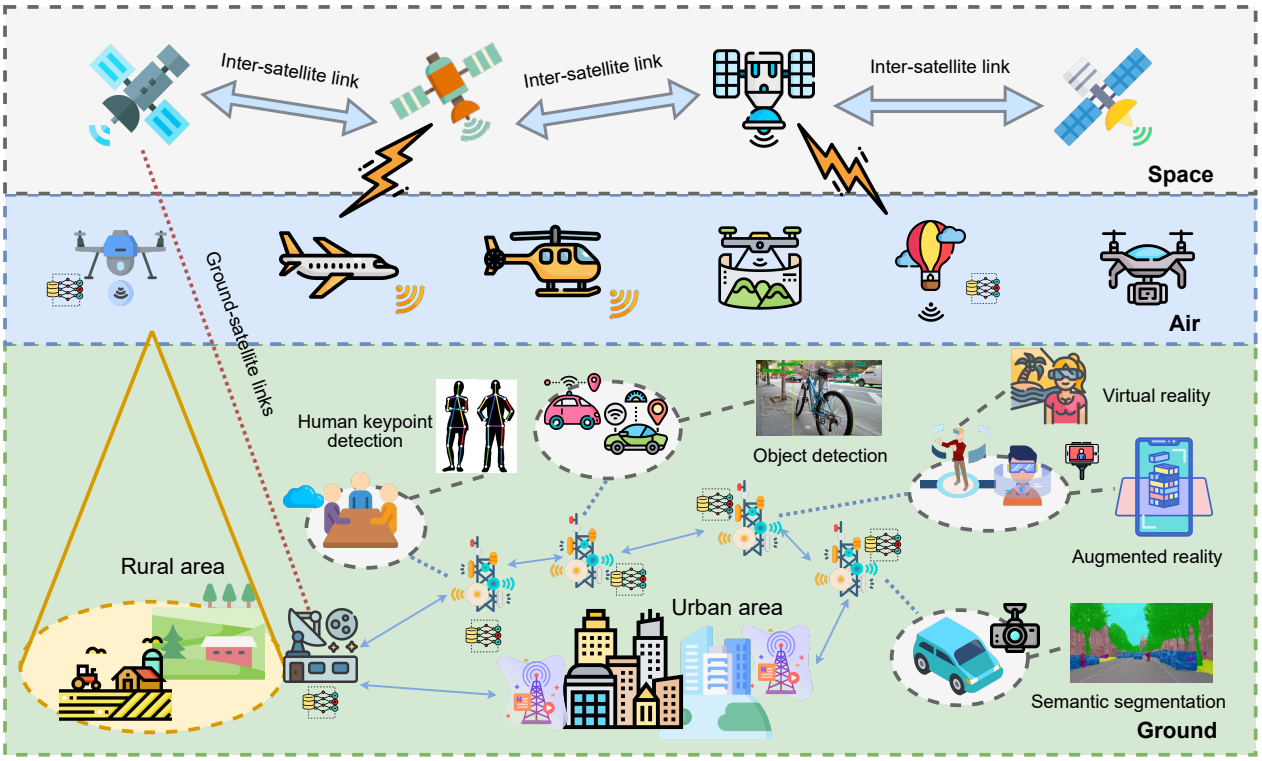


Fig. 4: A framework of 6G-enabled MVA for immersive XR.

The framework is divided into three layers: perception, network, and computation. The perception layer is composed of ubiquitous mobile devices with cameras, which continually capture videos from the physical world. The network layer comprises densely deployed terrestrial BSes, airborne BSes carried by flying vehicles (e.g., drones and airplanes), and satellite ground stations (GSes), which collectively provide seamless and high-quality network access for mobile devices worldwide. ECNs co-located with BSes constitute the computation layer to provide various MVA services.

A. Network- and Analytics-aware Compression

Image and video compression is necessary for MVA since uncompressed video frames are known to be very large and unsuitable for direct offloading. For example, an uncompressed YUV420 1080p frame has a size of 2.97 MB, let alone high-resolution 360° frames or volumetric frames including additional depth or point cloud information.

Traditional image and video compression algorithms developed for human visual systems, such as JPEG and H.264, are also popular in vision analytics systems. By controlling the compression parameters, one can trade off inference accuracy against offloaded data size (thus, the network traffic). For instance, Fig. 5 demonstrates how the image size and object detection accuracy vary with the JPEG image quality. As shown, with the image quality decreasing from 96 to 90, the data size reduces by 52.91% while the accuracy only drops by 1.17%. This suggests that 6G-enabled MVA should integrate compression parameters as configuration knobs to handle the unavoidable network dynamics and achieve the joint optimization of communication and computing.

Machine-centric image or video compression algorithms customized for specific vision tasks have recently gained increasing popularity, which typically achieve higher accuracy than the task-agnostic traditional compression algorithms [15]. They, however, usually end up with a prolonged E2E latency due to running DNNs for compression/decompression and poor generalizations for new tasks. Consequently, further efforts are needed to design novel, lightweight, and universal machine-centric image and video compression algorithms with minimum resource consumption. Additionally, the videos captured by a camera array, such as 360° and volumetric videos, are becoming popular due to their immersive nature. Yet, considering the particular geometry and the huge data volume, efficiently compressing them for MVA remains an open problem.

B. Dynamic Task Routing

We have seen the rapid deployment of 5G BSes recently. Till late 2022, China alone has deployed nearly 2.3 million BSes.[7] Since 6G would introduce tiny cells, the number would be even higher, providing better opportunities for ECN deployment. We expect 6G ECNs to handle various vision tasks and have mechanisms to select the most suitable model (e.g., YOLOv7 and EfficientDet[8]) for each task under certain

[7]https://www.rcrwireless.com/20230103/5g/chinese-telcos-deploy-2-million-5g-base-stations-nationwide [Accessed Mar 12, 2023]

[8]https://github.com/google/automl/tree/master/efficientdet [Accessed Mar 12, 2023]

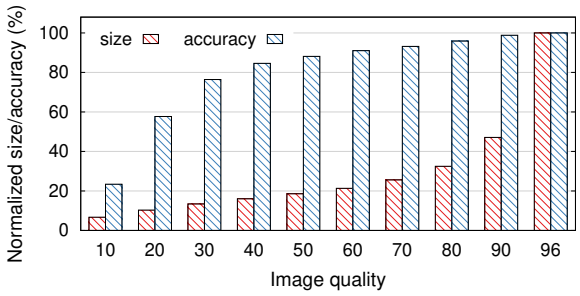


Fig. 5: The influences of image quality on image size and object detection accuracy. Results are normalized by that of the highest image quality. (model: YOLOv7; accuracy metric: mAP; image source: COCO's validation dataset).

resource constraints and performance requirements. However, the limited capacity of hardware resources (e.g., GPU memory) prevents ECNs from pre-loading a large pool of vision models simultaneously, while loading models on demand will introduce significant delays (up to tens of seconds).

With the dense deployment of 6G tiny cells, one mobile user may have multiple ECNs nearby. We thus suggest aggregating the computing resources of geographically proximate ECNs to improve the overall performance. A centralized controller can be deployed for distributed ECNs to guide the dynamic model loading/unloading on each ECN and route a user request to an appropriate ECN with a suitable model already being loaded. Each ECN can also exchange information with neighbor ECNs and make its model loading and task routing decisions if the centralized controller is unavailable.

The task routing among 6G ECNs also relies on accurate analytics workload prediction that accommodates the mobility of both users and ECNs. Resource management and scheduling techniques based on traditional workload predictions may no longer work. The trajectories of users and ECNs, the arrival patterns and performance goals of requests, the resource demands of vision models, and resource availability will be jointly considered to make task routing decisions.

C. Large-scale XR through Massive Collaborations

Current MVA tends to analyze video data from individual users isolatedly. The knowledge and insights gained are not shared between users, dramatically limiting one's perception of the physical world. In contrast, 6G-enabled MVA will develop the crowdsourcing intelligence to immerse everyone into a hybrid physical and virtual world. For instance, by jointly analyzing the massive video streams captured from different angles by audiences at a concert, complete awareness of the scene can be built in real time, allowing remote participants to interact with the scene from the comfort of their homes. Pervasive human-human and human-object interactions present another significant challenge in such a hybrid world. Multiple MVA tasks need to be executed simultaneously under the hood to deliver responsive and smooth interaction experiences. For instance, even simple interactions between an XR user and a virtual object require running object detection, hand tracking, and semantic segmentation in the background [12]. Therefore, resource-efficient multi-task MVA optimization is necessary. As users can also interact with objects in other ways, such as using voice or text prompts, combining MVA with analysis of other data modalities (e.g., audio and text) may enable more natural interaction.

The underlying MVA pipeline for large-scale collaborations may involve several stages, from denoising and synchronizing video data of massive cameras, analyzing them on distributed ECNs, to integrating distributed processing results for further analysis in cloud data centers. Ensuring a low E2E latency for real-time interaction is thus challenging. Fortunately, the accelerating vertical integration between MNOs and cloud providers (CPs) and horizontal integration between different MNOs or CPs have created new possibilities. For example, MNOs (e.g., Bell and Verizon) and CPs (e.g., AWS) have partnered to build 5G MEC with AWS Wavelength[9]. Additionally, sky computing[10] has emerged to accelerate the integration and simplify the use of multi-cloud resources. These techniques should undoubtedly be integrated with 6G-enabled MVA to realize immersive XR at scale.

Videos are privacy-sensitive data that may expose one's real-time locations, appearance, and property. Since shipping video data to ECNs is a network primitive in our framework, developing effective technologies to secure the offloaded data in transmission and computing is critical in 6G-enabled MVA. Advanced analytics tasks may also require federated analytics, where distributed ECNs collaboratively complete the analytics task without exchanging local data.

IV. Further Discussions

A. Integrating Terrestrial and Satellite Access Networks

To achieve a seamless global coverage, 6G is expected to integrate terrestrial and satellite access networks [8], offering services to remote and rural areas and bridging the digital divide. Worldwide XR services can then be realized, such as global conferencing, telemedicine and surgery, and wildlife surveillance. Although satellite access is already available in the latest smartphones, they generally use geosynchronous Earth orbit (GEO) or medium Earth orbit (MEO) satellites. Given the high orbits (35,786 km for GEO), these satellites have a wide coverage but suffer from long latency and narrow bandwidth. The services to mobile devices have then been largely confined to short messages for emergency use. For broadband and low latency access, 6G would eventually incorporate LEO (around 550 km) satellite communications to support real-time applications, such as XR.

Despite promising, LEO-supported MVA faces substantial challenges. For instance, each individual LEO satellite has a relatively limited coverage area. The serving LEO satellite for a mobile device can change in orders of minutes or even seconds. This requires efficient handover management, device-satellite relative movement predictions, and inter-satellite links to ensure service continuity. Current LEO constellations, such as Starlink, still rely on a *bent-pipe*, i.e., a satellite-to-GSes link to connect to the rest of the global Internet, which does not

[9]https://aws.amazon.com/wavelength/ [Accessed Mar 12, 2023]
[10]https://sky.cs.berkeley.edu/ [Accessed Mar 12, 2023]

fully unleash the potential of space communications. Cross-ocean inter-satellite links using laser remain in the experimental stage. Even if they are available, the round-trip latency will be no less than 60 ms across the Pacific Ocean, which is considerably lower than that of GEO satellite links (> 250 ms) or today's submarine fiber links (around 100 ms) but by no means comparable to accessing nearby BSes. As such, for a virtual scene constructed with worldwide distributed physical objects, smart synchronization is a mandate.

B. Energy-efficient and Energy-sustainable Analytics

DNN model inference can consume a significant portion of energy.[11] Therefore, energy-efficient vision models should be developed to avoid excessive energy consumption. The model inference is primarily executed in 6G ECNs, which can experience dramatic workload variations due to users' daily activities. Jointly coordinating distributed ECNs and designing workload-driven energy supply strategies will be essential to optimize network-wide energy efficiency. Moreover, mobile devices are prone to drain the embedded batteries, which can cause service interruptions. Designing an energy-efficient offloading strategy may mitigate this issue. Additionally, the energy transfer and harvesting components of 6G are hopeful of extending the battery life cycles. Mobile devices can even be made batteryless through backscatter communications.

MVA can produce large, bursty upload traffic, and transmitting the traffic to satellites may only work with high-end mobile devices with a robust energy supply. Our measurement shows that a current Starlink dish consumes over 60 W power for broadband communications at a speed of around 150 Mbps (DL) and 10 Mbps (UL). Relaying the traffic via airborne access networks established by drones or balloons could be promising and deserve further exploration.

C. Advanced Learning and Multimodal Analytics

We expect that federated learning will be a key to aggregating heterogeneous local data and resources of distributed ECNs for large model training. Meanwhile, transfer learning and meta-learning will be essential in reducing training costs and improving model adaptability. Beyond images captured from the physical world, vision models may be used to analyze artificial intelligence-generated content (AIGC) as it becomes pervasive and realistic. Adapting vision models for such content can be an interesting future direction. With multimodal data acquisition becoming affordable and reliable, multimodal analytics, which extends MVA from visual data to multimodal data sources (e.g., auditory and kinesthetic), deserves further exploration for better immersive experiences.

V. Conclusions

The full implementation of immersive XR relies on multiple technical enablers, such as intuitive human-machine interfaces. This article explores the communication and computing design optimization for MVA, a key enabling technology for XR. We started with a state-of-the-art review of MVA. We stressed that today's 5G design is not ready for massively distributed MVA, which struggles to balance latency and accuracy with constrained resources and fragmented operations. 6G would be a game changer; yet there are critical challenges ranging from ultra-fast network connections, ubiquitous access, to the seamless integration of computing and communication with smart task offloading. Our concrete vision of 6G-enabled MVA was described by an integrated framework empowered with edge intelligence. We also identified the critical issues towards its implementation and emerging enabling technologies.

[11] https://research.samsung.com/blog/Smart_at_what-cost_Characterising_Mobile_DNNs_in_the_wild [Accessed Mar 12, 2023]

Acknowledgment

This research was supported by a Canada NSERC Discovery Grant and a British Columbia Salmon Recovery and Innovation Fund (No. 2019-045).

References

[1] N. Wu, F. X. Lin, F. Qian, and B. Han, "Hybrid mobile vision for emerging applications," in *Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications (HotMobile'22)*, 2022, pp. 61–67.

[2] W. Zhang, Z. He, L. Liu, Z. Jia, Y. Liu, M. Gruteser, D. Raychaudhuri, and Y. Zhang, "Elf: Accelerate high-resolution mobile deep vision with content-aware parallel offloading," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom'21)*, 2021, pp. 201–214.

[3] M. Ghoshal, Z. J. Kong, Q. Xu, Z. Lu, S. Aggarwal, I. Khan, Y. Li, Y. C. Hu, and D. Koutsonikolas, "An in-depth study of uplink performance of 5g mmwave networks," in *Proceedings of the ACM SIGCOMM Workshop on 5G and Beyond Network Measurements, Modeling, and Use Cases*, 2022, pp. 29–35.

[4] X. Ran, H. Chen, X. Zhu, Z. Liu, and J. Chen, "Deepdecision: A mobile deep learning framework for edge video analytics," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1421–1429.

[5] L. Liu, H. Li, and M. Gruteser, "Edge assisted real-time object detection for mobile augmented reality," in *The 25th annual international conference on mobile computing and networking (Mobicom'19)*, 2019, pp. 1–16.

[6] L. Zhang, L. Chen, and J. Xu, "Autodidactic neurosurgeon: Collaborative deep inference for mobile edge intelligence via online learning," in *Proceedings of the Web Conference 2021 (WWW'21)*, 2021, pp. 3111–3123.

[7] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6g networks: Use cases and technologies," *IEEE Communications Magazine*, vol. 58, no. 3, pp. 55–61, 2020.

[8] F. Tariq, M. R. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6g," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 118–125, 2020.

[9] J. Meng, Z. J. Kong, Y. C. Hu, M. G. Choi, and D. Lal, "Do we need sophisticnsted system design for edge-assisted augmented reality?" in *Proceedings of the 5th International Workshop on Edge Systems, Analytics and Networking (EdgeSys'22)*, 2022, pp. 7–12.

[10] Y. Guan, X. Hou, N. Wu, B. Han, and T. Han, "Deepmix: Mobility-aware, lightweight, and hybrid 3d object detection for headsets," in *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys'22)*, 2022, pp. 28–41.

[11] P. Zhao, A. You, Y. Zhang, J. Liu, K. Bian, and Y. Tong, "Spherical criteria for fast and accurate 360° object detection," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI'20)*, 2020, pp. 12 959–12 966.

[12] D. G. Morín, P. Pérez, and A. G. Armada, "Toward the distributed implementation of immersive augmented reality architectures on 5g networks," *IEEE Communications Magazine*, vol. 60, no. 2, pp. 46–52, 2022.

[13] W. Saad, M. Bennis, and M. Chen, "A vision of 6g wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020.

[14] A. Narayanan, E. Ramadan, J. Carpenter, Q. Liu, Y. Liu, F. Qian, and Z.-L. Zhang, "A first look at commercial 5g performance on smartphones," in *Proceedings of The Web Conference 2020 (WWW'20)*, 2020, pp. 894–905.

[15] B. Chen, Z. Yan, and K. Nahrstedt, "Context-aware image compression optimization for visual analytics offloading," in *Proceedings of the 13th ACM Multimedia Systems Conference (MMSys'22)*, 2022, pp. 27–38.

BIOGRAPHIES

Miao Zhang (mza94@sfu.ca) received her B.Eng. degree from Sichuan University, Chengdu, China, in 2015, and her M.Eng. degree from Tsinghua University, Beijing, China, in 2018. She is currently a Ph.D. student in the School of Computing Science at Simon Fraser University, Burnaby, BC, Canada. Her research interests include cloud computing and multimedia systems.

Linfeng Shen (linfengs@sfu.ca) is currently a Ph.D. student in the School of Computing Science at Simon Fraser University, BC, Canada. He received B.Eng. in information security from Beijing University of Posts and Telecommunications in 2019, and M.Sc. in computing science from Simon Fraser University in 2021. His research interests include edge computing and multimedia.

Xiaoqiang Ma (mxqcs@ieee.org) received the B.Eng. degree from Huazhong University of Science and Technology, China, in 2010, and the M.Sc. and Ph.D. degrees from Simon Fraser University, Canada, in 2012 and 2015, respectively. He is currently a faculty member with the CSIS Department, Douglas College, Canada. His research interests include wireless networks, mobile computing, and cloud computing.

Jiangchuan Liu (S'01-M'03-SM'08-F'17) (jcliu@cs.sfu.ca) is a Professor at Simon Fraser University, BC, Canada. He is a Fellow of The Canadian Academy of Engineering and an IEEE Fellow. He received B.Eng. (cum laude) from Tsinghua University and Ph.D. from HKUST. He has served on the editorial boards of IEEE/ACM Transactions on Networking, IEEE Transactions on Multimedia, IEEE Communications Surveys and Tutorials, and IEEE Internet of Things Journal. He was a Steering Committee member of IEEE Transactions on Mobile Computing and Steering Committee Chair of IEEE/ACM IWQoS. He was TPC Co-Chair of IEEE INFOCOM'2021.