Edge Intelligence Empowered Immersive Media: Challenges and Approaches

Zhi Wang
Tsinghua University

Jiangchuan Liu
Simon Fraser University

Wenwu Zhu
Tsinghua University

***Abstract*—**
Recent years have witnessed many immersive media services and applications, ranging from 360-degree video streaming, to augmented and virtual reality (AR and VR) and the recent metaverse experiences. These new applications usually have common features including high fidelity, immersive interaction, and open data exchange between people and the environment. As an emerging paradigm, edge computing has become increasingly ready to support these features. We first show that a key to unleashing the power of edge computing for immersive multimedia is handling AI models and data. Then, we present a framework that enables joint accuracy- and latency-aware edge intelligence, with adaptive deep learning model deployment and data streaming. We show that not only conventional mechanisms such as content placement and rate adaptation, but also the emerging 360-degree and virtual reality streaming can benefit from such edge intelligence.

Introduction

Recent years have witnessed many immersive multimedia services and applications, including 360-degree video streaming, augmented and virtual reality, and the current multi-interface supported metaverse applications [1], [2]. These new applications usually have common features: high fidelity (*e.g.*, 4K resolution, 90+ fps, and 100+ degrees of field of view are supposed to support satisfactory Quality of Experience (QoE)), immersive interaction (*e.g.*, $<$ 10 milliseconds in interaction latency), and open data exchange (*e.g.*, any users among the billions of today's social-network users can communicate in virtual worlds to share all kinds of multimedia content via different social groups).

These features require powerful computational resources in the service infrastructure, low latency in the network, and social network support. Today, solutions to realize these immersive multimedia applications are mainly either cloud-centric or device-centric. In cloud-based solutions, the centralized cloud services (*e.g.*, virtual GPU/CPU instances, databases, and content delivery networks) are utilized. Multimedia content, such as video chunks, are streamed to users from geo-distributed servers deployed worldwide. For device-based solutions, more content processing workflows are carried out by users' end devices like head-mounted displays. Such devices are usually

Published by the IEEE Computer Society

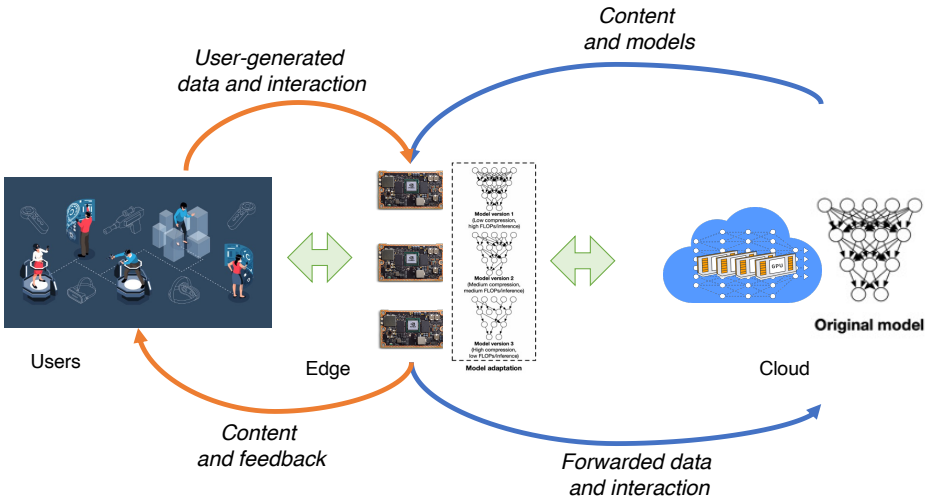

Figure 1. Illustration of the edge intelligence for immersive multimedia applications.

powered with local computational capacities and storage so that a certain fraction of the applications are running at the local devices, *e.g.*, virtual reality scenes can be rendered and displayed locally.

Although these two paradigms have many implementation feasibilities and infrastructure readiness, they have the following limitations for high-quality immersive multimedia applications. On one hand, cloud-based solutions assume that all or most of the multimedia functionalities can be deployed remotely, *i.e.*, in the cloud, which is not feasible for services and applications with stringent data privacy requirements [3]; Meanwhile, due to the intrinsic propagation latencies between a cloud server and the end-users, and latencies caused by cloud infrastructure (*e.g.*, load balancer and other redirection mechanisms), certain low latency is hard to be achieved in cloud-based solutions, especially when users are located far away from the cloud servers. At the same time, today's cloud deployment is also more expensive in general due to the installation and maintenance costs for dedicated infrastructure. On the other hand, the biggest limitation of a device-centric solution is the limited energy, computational, and storage resources on a user-end device, making it hard to provide a quality guarantee for emerging resource-intensive multimedia applications; Also, device-centric solutions usually assume a standalone installation on a particular device, making it less flexible to the application evolution.

Due to these limitations, new paradigms have been investigated. Among these attempts, using edge computing to empower immersive multimedia has become promising in recent years [4], [5], [6], [7]. The infrastructure is to be accessible at the edge for the immersive metaverse, providing functionalities including communication and networking, computation, and also blockchain [8]. As an emerging infrastructure design paradigm, edge computing is usually deployed much closer to users as a "middle layer" between the cloud and the user devices. As illustrated in Figure 1, in edge scenario, on one hand, interactions and data generated by users in immersive multimedia applications can be processed by the edge devices, in a "device-to-edge" offloading manner; On the other hand, cloud can also offload certain tasks to the edge, *e.g.*, letting edge execute compressed deep learning models to provide *inference* services locally. Edge computing has become increasingly ready to support the above-required features, with much lower cost and latency, high scalability, and better privacy protection. Compared with device-centric solutions, edge computing is more resourceful and scalable,

and easier to adapt to application deployment over time. According to State of The Edge report[1], the global IT power footprint for edge infrastructure is forecasted to increase from 1 GW in 2019 to over 40 GW by 2028, and 37% of the edge infrastructure will be used by mobile and residential consumers, the major users of advanced immersive multimedia services and applications.

Our study serves as an exploration to powering edge computing for enabling better immersive multimedia. We first show that the features in immersive multimedia services and applications, can be well supported by edge computing. Next, we show that *QoS (Quality-of-Service)* in immersive multimedia includes both traditional metrics such as bitrate and streaming delay, and the new metrics related to learning-based strategies such as inference accuracy and latency. The key to unleashing the power of edge intelligence, is to empower the edge devices to handle AI models and user data well, with QoS-guaranteed inference latency and accuracy.

We propose a framework that enables accuracy- and latency-aware edge intelligence, with adaptive deep learning model deployment and data streaming. Several learning-based strategies have shown their capabilities in edge scenarios as follows: i) Predictive models, which have been designed to predict resource demands and user experience in immersive media; ii) Deep Reinforcement Learning (DRL), which is a combination of Deep Learning (DL) and Reinforcement Learning (RL) to build an agent to learn the best actions over a set of states through the interaction with the environment, so as to maximize the long-term accumulated rewards; and iii) Meta-learning, which is an array of methodologies employed to optimize the performance of existing deep learning models, *e.g.*, for the changing edge data distributions. Edge intelligence powered by such learning methods can help strategies in immersive multimedia services and applications, including content placement and rate adaptation, as well as viewport prediction in 360-degree and VR streaming.

[1]https://stateoftheedge.com/reports/state-of-the-edge-2020/

Edge Intelligence for Immersive Multimedia

As illustrated in Figure 2, for the high fidelity feature that usually requires a QoS-guaranteed network, edge topology and content caching are promising to deliver multimedia content. For the immersive user interactions, edge infrastructure allows part of the interactions to be responded by modulars executing locally. And for the open data exchange, by mechanisms such as *simulating* a virtual "avatar" locally on edge devices, communicating with remote proxies can possibly be avoided in some cases.

- High fidelity means a high-quality image/audio experience, and it usually requires that the multimedia content can be delivered to users with high QoS and QoE. In today's multimedia service distribution paradigms, research efforts have been mainly devoted into two areas: 1) Building high-quality service-to-user network topology and infrastructure, so that multimedia content and data can be streamed to users with end-to-end bandwidths large enough; 2) Building efficient data prefetching and caching strategies to enable users to fetch the needed content from local.
- Immersive user interaction is also an essential feature in immersive multimedia applications. According to different system topology and communication patterns, the response quality of user interaction depends on the following factors. 1) *Local processing*, where the interaction response can be processed locally, *e.g.*, game scenes that are already pre-downloaded and the user's interaction feedback can be generated locally; and 2) *Remote processing*, where the interactions are to be propagated to remote servers or other users, and the user has to receive certain information correspondingly to generate local feedback. In either case, the system will process specific tasks on the edge devices to reduce the remote communication overhead.
- Open data exchange is another feature of immersive multimedia applications. Among

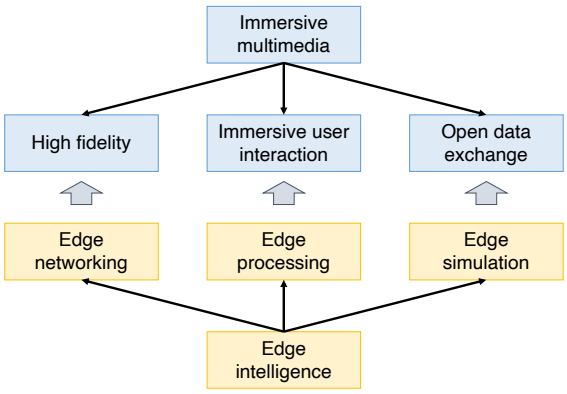

Figure 2. Edge intelligence for supporting immersive multimedia applications and services.

the previous immersive interactions, some interactions involve data exchange between people. In immersive applications such as virtual events or gaming, the users can be physically located in different places in the world, and open data exchange is challenging because messages have to be delivered for a long distance, incurring large latencies. Handling such exchange locally is a promising strategy to reduce communication latency. Since such response is performed by users, edge intelligence provides the potential to "generate" possible responses *similar* to those generated by real remote users according to their historical behaviors, so that some messages to a remote user can be handled locally.

Key to Edge Intelligence: Handling Model and Data Well

The key to edge intelligence is about how to handle deep learning models well. In particular, handling models well means satisfying users with the expected accuracy and latency during the inference phase [9], after the deep learning models have already been pre-trained for deployment. It usually involves collaborative model inference with both edge and cloud capacities, *e.g.*, joint edge-cloud inference.

Different from conventional cloud-based MLaaS (*i.e.*, Machine Learning as a Service) in which inference accuracy is regarded as the most if not only criteria for the inference performance, applications in immersive multimedia usually have multiple inference performance metrics varying over time, including 1) inference latency that determines the "liveness", 2) inference complexity that determines the energy consumption, and also 3) inference accuracy.

Impact of Model and Data on Inference Quality

In the edge inference setup for immersive multimedia, these metrics are actually affected by both the deep learning model (usually "customized" to fit the edge devices) and the input data (usually "compressed" to fit the network condition) for the inference tasks. Next, we present how model and data affect the inference quality, respectively.

Impact of Model on Inference Quality When using edge intelligence in immersive multimedia applications, deep learning models are executed entirely or partially at the edge devices. The impact of a deep learning model on inference performance is highly affected by the model structure and training status. Keeping the model deep and large enough is a common practice to have high accuracy, but it also incurs high computational costs for each inference task, leading to larger inference latency. Knowledge distillation, model pruning, and model quantization have thus been used to make "lightweight" model versions [10], which particularly satisfy resource-limited edge devices.

Impact of Data on Inference Quality On the other hand, in the edge intelligence for the immersive multimedia scenario, input data is also a non-negligible factor to affect the inference quality, in a sense that the input data (*e.g.*, video stream, image, and audio) has to be uploaded or streamed from user devices to edge devices and/or cloud servers. The input data thus determines the upload latency by its volume and the inference accuracy by its data quality (*e.g.*, image resolution and the inherent inference difficulty). Data compression strategies (*e.g.*, JPEG) and downsampling (*e.g.*, region crop and pixel-level downsampling) are common practices to change the input data, which all lead to the

change in inference quality.

In our study, we have particularly investigated the impact of both the lightweight model versions and the different compressions of the input data on the inference performance. In Figure 3, we plot the impact of model pruning and quantization on the inference accuracy: the x-axis represents the different data versions (larger x indicates smaller input data size), and the y-axis represents the different model versions (larger y indicates smaller model complexity and size). Each curve represents the same accuracy achieved with different model and data configurations. In Figure 3(a)(b), we have the impact of varying model pruning and quantization with input images compressed with different JPEG quality levels; While in Figure 3(c)(d), we have the impact of varying model pruning and quantization with input images compressed by different downsamplings. Our insights for changing models and input data for edge intelligence are as follows.

- Larger models provide higher accuracy consistently under different data compression schemes, and larger data sizes also ensure higher accuracy, under different model compressions.
- The same inference accuracy can be achieved by different configurations of model versions and data versions, as long as they are in the same accuracy curve in these figures. This suggests that we are able to deploy models strategically to serve users because we can balance the tradeoff between the size for data transmission and the computational complexity for model execution in the inference.

To empower edge infrastructure for immersive multimedia services and applications, a common demand is to deploy models or partial models on edge devices, whose intelligence can be used for optimizing the edge network topology, content processing as well as avatar simulation. A challenging problem for edge deep learning deployment is how to balance the tradeoff between the accuracy and inference latency, both affected by the complexity of deep learning models.

Privacy in Edge Intelligence for Immersive Multimedia

In today's practice, edge intelligence is intrinsically enabled by a data-driven scheme. Thus, privacy and security are non-negligible issues for edge intelligence-driven immersive multimedia, especially when personal data can be collected from more and more new modalities. Looking into the future metaverse applications, security and privacy issues are likely to become even more important concerns [3], *e.g.*, pervasive data collection, privacy leakage, and compromised edge/end devices.

In the edge intelligence infrastructure, privacy and security can be enhanced by computational offloading using a trusted execution environment, federated learning, and also adversarial machine learning [8]. In the edge intelligence framework, the training process can use federated learning and adversarial machine learning for privacy protection, while the inference process can benefit from trusted execution environment. Meanwhile, it is also important to leverage machine learning and blockchain technologies to promote self-governance capabilities of metaverse communities to improve privacy and security, since intelligence and distribution are two trends for immersive and future metaverse regulation.

Joint Accuracy- and Latency-aware Model Deployment

To deploy models at edge devices to satisfy different accuracy and latency requirements of users, we have the following model deployment framework. Figure 4 illustrates the accuracy- and latency-aware deep learning model edge deployment. We propose to "profile" the impact of both deep learning models and input data samples, on the inference accuracy and inference latency. In particular, the purpose of profiling a deep learning model is to compress models so that they can be executed on such resource-limited edge devices, and the purpose of data profiling is to compress input data so that they can be transmitted fast enough. Since model and data compressions are not free—they usually lead to degraded accuracy, strategical model compression and deployment, as well as data compression and

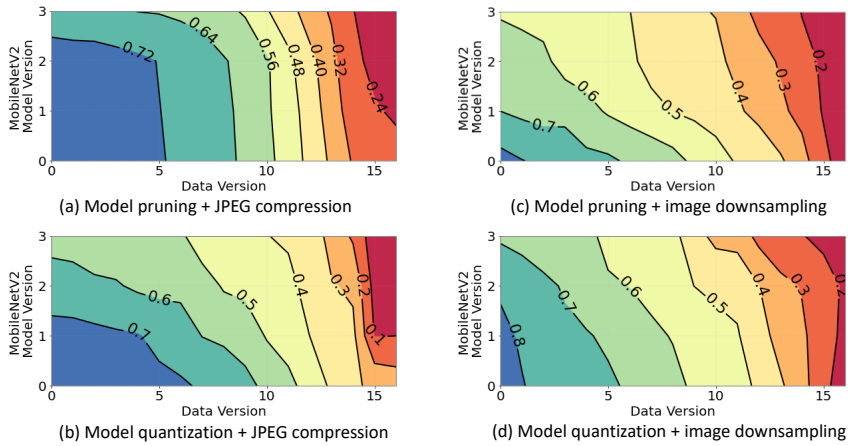

Figure 3. The impact of model versions and data compressions on the inference accuracy.

inference task scheduling, are in demand.

Profiling Model and Data

Model and data profiling is the first step for model deployment and data scheduling. For today's sophisticated deep structures like Transformer, it is challenging to have a closed-form analysis for the impact of deep learning models and quality of input data on the final inference accuracies. We propose a data-driven approach for profiling models and data on their impact on the inference performance. We have an original model M and a series of model compression "operations" (*e.g.*, pruning, quantization, etc.), which compress the original model into its variants $\{M_1, M_2, \cdots, M_m\}$, where M_i is the i-th compressed version of model M. We have a testing dataset D, which can be divided into subsets $\{D_1, D_2, \cdots, D_d\}$, where each subset D_k represents certain features like "daytime", or "outdoor", etc.. Similarly, a data sample $d \in D_k$ can be compressed by a data compression operation, such as JPEG or downsampling for image data, into different versions $\{d_1, d_2, \cdots, d_s\}$, where d_j is the j-th compression version for sample d.

We calculate a relative accuracy for the compressed model and data, against the accuracy inferring the original data sample over the original large model. When "profiling" the performance for a particular combination of a pair of model compression strategy generating a model compression, and a data compression strategy generating data compressions for a particular sub-dataset D_k, we calculate the average relative accuracy, so that it captures how the model compression and data compression perform for samples similar to D_k.

In our study, we have observed that for different subsets, the "best" combinations of model compression and data compression strategies also differ. When running the compressed model on a particular edge device, besides the accuracy above, we also consider the inference latency, *i.e.*, the time spent on inferring each input sample or sample batch. In the profiling, we measure the inference latency consisting of the upload delay and model execution delay, when executing an input sample over a compressed model over an edge device. Iteratively, we are able to estimate the inference accuracy and latency for all combinations of model compressions and data compressions, for all the data distributions selected for the profiling.

Joint Model Deployment and Data Sampling

The tradeoff between inference accuracy and latency allows us to use different combinations of model compressions and data compressions to yield similar accuracy and latency metrics, which provides a new design space for a joint model deployment, and data sampling

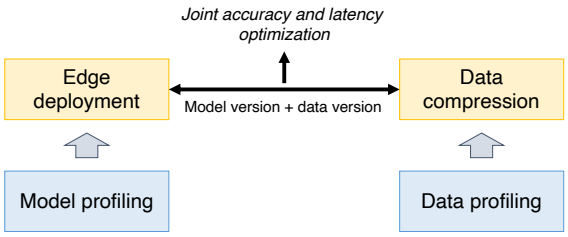

Figure 4. Framework of accuracy- and latency-aware deep learning model edge deployment.

and request scheduling.

So, we propose to jointly generate different model versions and compressed data versions and schedule the input samples under different compression levels to different model servers. As such scheduling is based on the accuracy and latency profiling above, it can satisfy diverse requirements in use cases, with different accuracy and latency requirements, *i.e.*, different accuracy + latency combinations, with minimized model serving costs. In practical immersive multimedia applications, the deep learning models of different versions can be deployed across geo-distributed edge devices so as to receive inference queries accordingly. In particular, we formulate the compressed model deployment and input data sample scheduling as an optimization program to minimize the overall costs.

Joint multi-version model deployment and input data compression and scheduling allow deep learning inference to power edge infrastructure to support immersive multimedia applications. Next, we present how such edge intelligence is able to improve multimedia content replication/caching and learning-based rate adaptation, and viewport prediction in 360-degree video streaming.

Case Studies for Edge Intelligence Powered Immersive Media

In this section, we show how conventional strategies such as content placement and rate adaptation, can benefit from edge intelligence, with deep learning and inference from edge infrastructure; We also present strategies like viewport prediction which is important in immersive multimedia applications, can be enhanced by edge intelligence.

Edge Caching for Multimedia Content

Content placement deploys content close to users so as to improve the availability of content nearby, to improve quality of user experience. In emerging multimedia applications, "small" content providers (CPs) play an important role due to the intrinsic diverse preferences of users [11]. A major difference between small CPs and traditional CPs is that small CPs usually allocate resources from both cloud and edge content delivery networks, in a pay-as-you-go manner, instead of building their own infrastructures. Such "content delivery as a service" with edge-network configurations allows many flexible configurations, *e.g.*, *scaling cache size* and *changing the content replacement strategies* for more dynamic strategies. One potential technical challenge is that jointly scaling the cache capacity and changing the content placement strategies require more sophisticated edge intelligence, as compared to traditional single-dimension decision.

Edge Learning-based Caching Agent. We propose an edge-based joint caching strategy. In particular, we design a reinforcement learning model at edge caches, to jointly and dynamically tune the edge cache size and replacement strategy, with the following major characteristics.

- *Working for dynamical popularity changes.* Small CPs of immersive multimedia applications usually have dynamical content deployment patterns, challenging the learning process of the RL agent. We keep a dynamically-changing learning window of historical requests, so that the agent can learn from both *stationary* and *non-stationary* popularity distributions.
- *Supporting different decision spaces.* Cache capacity decision is usually a continuous action, and content placement strategy is usually a categorical variable. To alleviate performance degradation that existing deep reinforcement learning algorithms suffer when there are joint continuous and categorial actions, an action "fusing" scheme can be utilized to fuse the different action spaces, so that they can be adjusted to achieve a joint performance gain.

- *Deployable at edge devices for fast inference.* To allow fast deployment and inference at edge devices, we use the model and data compressions to enable fast content caching strategies. In trace-driven experiments with dynamical user requests, the edge learning-based caching is able to improve the cache hit rate by over 20% against conventional size-only solutions, and over 3% against conventional strategy-only solutions, both with reduced deployment costs [11].

Learning-based Rate Adaption

Rate adaptation is also an important strategy in immersive multimedia that benefits from learning-based schemes. In recent years, learning-based strategies have been recognized as an effective means to overcome unnecessary playback rebuffering and network fluctuation. These strategies do not rely on pre-programmed models/assumptions about the environment and gradually learn the policies for bitrate decisions through observation [12].

Maximizing the quality of experience for users has been the major purpose for traditional rate adaptations such as Pensieve [13]. These conventional strategies rely on the information about video chunk sizes to guide the adaptation decisions, of which a most important prerequisite is future video chunk size. However, in immersive multimedia applications with live streaming characteristics, such information may not be available all the time, due to the changing content generated real-time. For example, in scenarios like crowd-sourced live streaming, the video chunks are generated and transcoded in real time so the future video chunk size is unknown ahead.

Chunk Size Prediction with Meta Learning Intuitively, one can use the average of the past video chunk sizes as the estimated future video chunk size, but this leads to large errors for videos with more dynamical bitrate. Deep Neural Network (DNN)-based models have then been designed to predict video chunk size [14]. For different videos, meta-learning is an effective way to train similar tiny models fast [15], for video sessions with different content characteristics. Based on the predictions, reinforcement learning-based agents then use the prediction as state inputs and QoE metrics as rewards, to adjust future bitrate decisions for rate adaptation.

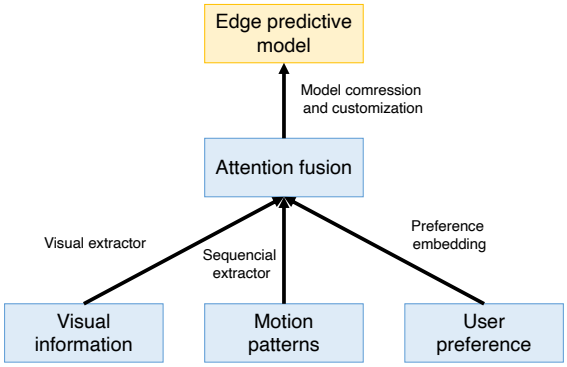

Figure 5. An edge viewport predictive framework: a spherical CNN based 360-degree video feature extraction network for multi-modular information fusion, and model compression for edge.

Edge Intelligence for RL Adaptation We propose an edge learning-based rate adaptation framework for immersive live streaming, by extracting information from videos to predict the bitrate of the video streaming.

- First, we propose a DNN-based bitrate prediction network of future video chunks using video coding meta information (*e.g.*, the residual frames, quantization parameters, and the division information of macroblocks in H.264). We pass the meta information in the encoding process to the model, and use meta-learning to train the chunk size prediction network.
- Second, we use deep reinforcement learning to design a live adaptation framework based on the bitrate predictions. To achieve low latency for immersive applications, the meta learning prediction framework and RL-based rate adaptation agent are deployed locally on edge devices to make the rate selection decisions.

Viewport Prediction for 360-degree and VR Streaming

As an important immersive multimedia application, 360-degree video streaming has become popular in various video streaming

platforms. To improve users' quality of experience and engagement, using viewport prediction has become a *de facto* practice for content prefetching, and the previous common practice is to utilize multi-modal information, including user behaviors, content characteristics and the recent viewing patterns, to improve prediction accuracy [16]. Due to the high stringent latency requirement and the multi-modular information needed for viewport prediction, deploying the prediction models close to users on the edge devices is a promising solution.

Multi-modular Predictive Network In a 360-degree video view session, a user's viewport changing is not only correlated with the content information but also other information including the head motion behaviors of the user as well as his/her intrinsic preferences. We propose to jointly use users' recent viewing patterns, the visual information, as well as the motion and preference information to build a viewport prediction model as follows.

- Although spherical convolutional neural networks (CNNs) [17] show certain potential in processing spherical data, they are still at an early age and cannot handle several real-world situations, *e.g.*, they cannot achieve rotation-invariant convolution since traditional CNN only has the ability of translational-invariant convolution. We propose a dedicated spherical CNN based 360-degree video feature extraction network. As illustrated in Figure 5, it takes in multi-modular information, including viewing behaviors, the video content, and the head motion patterns.
- We use the user's personal preference as a context embedding, which serves as the spatial attention to the video content, *i.e.*, a semantic "relevance" between a particular region on future video and the user's preference. In particular, we propose to use a convolutional recurrent neural network (RNN)-based feature extraction model to capture the joint spatial and temporal characteristics from the sequentialized 360-degree video frames. It encodes the user preference by referring to the videos a user has viewed before, with the RNN model to incorporate viewing histories so that the model is expected to represent viewport changes in the future time window. The proposed model with near-identity kernel is able to achieve an accuracy of 65.7%, which outperforms a model with equatorial kernel [16].

Edge-based Viewport Prediction After the multi-modular model is customized and trained for a particular user, we deploy it at an edge device close to the user for fast inference, for the purpose to guide content fetching or other interactions, as illustrated in Figure 5. The model again is compressed using the previous practice, according to the expected accuracy and latency requirement with the constraint of the edge device computational capacities.

In Figure 6, we plot the prediction visualization for two selected videos from dataset UCF101[2], which is a representative dataset of human actions, consisting of realistic user-uploaded videos with camera motion and cluttered background. We use w_k to denote the future time window (in second) for the viewport prediction, and a larger w_k usually requires a more powerful prediction capacity.

In this figure, the viewports in the red rectangles are the ground truth, *i.e.*, real viewports captured in the datasets, and the blurred viewports are the predictions. For a large fraction of the cases, the predictive results are relatively consistent with the real viewports. In the "successful" predictions (*e.g.*, 1st, 2nd, 4th in the first row and 3rd and 4th in the second row), we observe that even with rapid head movements at $w_k = 1$ and $w_k = 2$, our model is still able to perform properly. In the "failure" cases (*e.g.*, 3rd in the first row and 1st and 2nd in the second row), the predictive network fails at different head movement speeds. These findings suggest that the prediction capability of the design is not entirely driven by the head movement intensity, but also affected by the visual content

[2] https://www.crcv.ucf.edu/data/UCF101.php

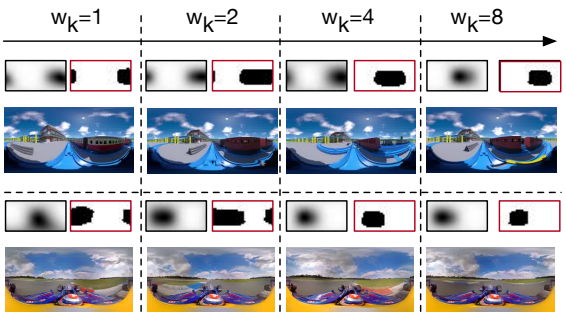

Figure 6. Performance visualization for the edge viewport prediction.

itself.

Concluding Remarks

Immersive multimedia services and applications are emerging, either evolving from traditional crowdsourced live streaming and 360-degree video streaming, or growing from augmented and virtual reality applications. We discussed how some common requirements can be satisfied by edge intelligence, with much lower cost and latency and better privacy protection. To unleash the power of edge computing for immersive multimedia, we provided a joint model compression and data scheduling framework that enables accuracy- and latency-aware edge intelligence, with adaptive deep learning model deployment and data streaming. We also showed that edge content replication and learning-based rate adaptation are readily supported by such edge intelligence. Furthermore, we used a 360-degree video streaming case to demonstrate that edge-assisted viewport prediction can be achieved and used for tile prefetching, leading to improved quality of user experience in similar immersive-media applications.

We have seen exciting development toward edge intelligence and immersive multimedia separately and jointly. With the evolution of today's multimedia technical stack, we are seeing new trends for immersive multimedia, including new video coding such as Versatile Video Coding (VVC) for 360-degree and VR video streaming, and the next-generation advanced mobile communications system, 6G, which is supposed to provide communication links to fuse the physical, cyber, and biological worlds, giving rise to immersive applications with new possibilities enabled. We believe there are tremendous imagination and research opportunities, to realize these visions, with the hope of making truly high-fidelity, immersive-interaction, social, and also personalized immersive multimedia a reality.

Acknowledgments

This work was supported in part by Shenzhen Science and Technology Program (Grant No. RCYX20200714114523079).

REFERENCES

1. C.-L. Fan, W.-C. Lo, Y.-T. Pai, and C.-H. Hsu, "A survey on 360 video streaming: Acquisition, transmission, and display," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–36, 2019.
2. H. Duan, J. Li, S. Fan, Z. Lin, X. Wu, and W. Cai, "Metaverse for social good: A university campus prototype," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 153–161.
3. Y. Wang, Z. Su, N. Zhang, D. Liu, R. Xing, T. H. Luan, and X. Shen, "A survey on metaverse: Fundamentals, security, and privacy," *arXiv preprint arXiv:2203.02662*, 2022.
4. C. Long, Y. Cao, T. Jiang, and Q. Zhang, "Edge computing framework for cooperative video processing in multimedia iot systems," *IEEE Transactions on Multimedia*, vol. 20, no. 5, pp. 1126–1139, 2017.
5. J. Martín-Pérez, L. Cominardi, C. J. Bernardos, A. de la Oliva, and A. Azcorra, "Modeling mobile edge computing deployments for low latency multimedia services," *IEEE Transactions on Broadcasting*, vol. 65, no. 2, pp. 464–474, 2019.
6. P. Roy, S. Sarker, M. A. Razzaque, M. M. Hassan, S. A. AlQahtani, G. Aloi, and G. Fortino, "Ai-enabled mobile multimedia service instance placement scheme in mobile edge computing," *Computer Networks*, vol. 182, p. 107573, 2020.
7. F. Alvarez, D. Breitgand, D. Griffin, P. Andriani, S. Rizou, N. Zioulis, F. Moscatelli, J. Serrano, M. Keltsch, P. Trakadas *et al.*, "An edge-to-cloud virtualized multimedia service platform for 5g networks," *IEEE Transactions on Broadcasting*, vol. 65, no. 2, pp. 369–380, 2019.
8. M. Xu, W. C. Ng, W. Y. B. Lim, J. Kang, Z. Xiong, D. Niyato, Q. Yang, X. S. Shen, and C. Miao, "A full dive into realizing the edge-enabled metaverse: Visions,

enabling technologies, and challenges," *arXiv preprint arXiv:2203.05471*, 2022.

9. J. Jiang, Z. Luo, C. Hu, Z. He, Z. Wang, S. Xia, and C. Wu, "Joint model and data adaptation for cloud inference serving," in *2021 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, 2021, pp. 279–289.
10. A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," *arXiv preprint arXiv:1802.05668*, 2018.
11. J. Ye, Z. Li, Z. Wang, Z. Zheng, H. Hu, and W. Zhu, "Joint cache size scaling and replacement adaptation for small content providers," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.
12. A. Bentaleb, B. Taani, A. C. Begen, C. Timmerer, and R. Zimmermann, "A survey on bitrate adaptation schemes for streaming media over http," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 562–585, 2018.
13. H. Mao, R. Netravali, and M. Alizadeh, "Neural adaptive video streaming with pensieve," in *Proceedings of the conference of the ACM special interest group on data communication*, 2017, pp. 197–210.
14. A. Lekharu, S. Kumar, A. Sur, and A. Sarkar, "A qoe aware lstm based bit-rate prediction model for dash video," in *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*. IEEE, 2018, pp. 392–395.
15. R. Vilalta and Y. Drissi, "A perspective view and survey of meta-learning," *Artificial intelligence review*, vol. 18, no. 2, pp. 77–95, 2002.
16. C. Wu, R. Zhang, Z. Wang, and L. Sun, "A spherical convolution approach for learning long term viewport prediction in 360 immersive video," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 14 003–14 040.
17. T. S. Cohen, M. Geiger, J. Köhler, and M. Welling, "Spherical cnns," *arXiv preprint arXiv:1801.10130*, 2018.

Zhi Wang (S'10-M'14-SM'22) is currently an associate professor at Shenzhen International Graduate School, Tsinghua University. He received his Ph.D. in 2014 and his B.E. in 2008, both from Tsinghua University. His research areas include multimedia networks, mobile cloud computing, and large-scale machine learning systems. His research won the Best Paper Award of ACM Multimedia, the Best Student Paper Award of MMM, and the Best Paper Award of ACM Multimedia, HUMA Workshop. He is an Associate Editor of IEEE TMM. Contact him at wang_zhi@tsinghua.edu.cn.

Jiangchuan Liu (S'01-M'03-SM'08-F'17) is a University Professor in the School of Computing Science, Simon Fraser University, British Columbia, Canada. He is a Fellow of The Canadian Academy of Engineering, an IEEE Fellow, and an NSERC E.W.R. Steacie Memorial Fellow. He was an EMC-Endowed Visiting Chair Professor of Tsinghua University (2013-2016). In the past he worked as an Assistant Professor at The Chinese University of Hong Kong and as a research fellow at Microsoft Research Asia. He received the BEng degree (cum laude) from Tsinghua University, Beijing, China, in 1999, and the PhD degree from The Hong Kong University of Science and Technology in 2003, both in computer science. Contact him at jcliu@sfu.ca.

Wenwu Zhu (S'91–M'96–SM'01–F'10) is a Professor in Department of Computer Science and Technology, Tsinghua University. He is an AAAS Fellow, IEEE Fellow, SPIEFellow, and a member of the Academy of Europe. Prior to his current position, he was a Research Manager at Microsoft Research Asia, Beijing, China. He was the Chief Scientist and Director with Intel Research China, Beijing,China, from 2004 to 2008. He was at Bell Labs, Murray Hill, NJ, USA, as a member of technical staff from 1996 to 1999. He received the Ph.D. degree from NewYork University, New York, NY, USA, in 1996. Contact him at wwzhu@tsinghua.edu.cn.