Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5565648, 21 pages
https://doi.org/10.1155/2021/5565648

Review Article

A Survey on Caching in Mobile Edge Computing

Yuhan Zhao[1], **Wei Zhang**[1,2], **Longquan Zhou**[3] **and Wenpeng Cao**[1]

[1]*Shandong Provincial Key Laboratory of Computer Networks, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan 250000, China*
[2]*State Key Laboratory of High-End Server & Storage Technology, Jinan 250000, China*
[3]*Department of Information Engineering, Shandong University of Science and Technology, Tai'an 251000, China*

Correspondence should be addressed to Wei Zhang; wzhang@sdas.org

Received 17 February 2021; Accepted 30 September 2021; Published 10 November 2021

Academic Editor: Bithas Petros

Boosted by the visions of 5G technology, along with the proliferation of IoT devices, we are lucky to bear witness of a tremendous shift within the Internet, mobile computing, and ubiquitous applications pervading people's daily life. To catch up with the speed of this evolution, new architectures, which enable us to decentralize and focus more on the edge of the network, have to show up. To effectively address the record-breaking growth of data traffic, we also have to come up with new caching mechanisms in coping with users' demanding QoE along with other aspects, such as data privacy and energy efficiency. In this paper, we intend to begin with reviews on edge caching. First, we make a detailed overview of mobile edge caching. Then, we move on to the QoS and QoE part, discussing the related literature concerning these topics. We then start to address factors concerning edge caching and QoE. Four cutting-edge applications utilizing these technologies with challenges posed by current network architectures are made clear. Our paper closes with some future research directions.

1. Introduction

Fast-growing development of the Internet along with the ubiquitous IoT devices has led to a skyrocketing growth of traffic via mobile communications over the past decade. Meanwhile, massive IoT terminal devices are playing an increasingly indispensable role in agriculture, medical care, education, transportation, modern economy, new energy, smart home, environment monitoring, and other industries [1]. The Cisco Visual Networking Index 2018 [2] made an estimation that global mobile Internet traffic is about to increase sevenfold with a composite annual growth rate of 47 percent between 2016 and 2021. And with statistics provided by Tencent Cloud service, mobile network traffic took up to 98.7 percent of the whole network in 2018, while it seems to be on a much slower but steady growth. Likewise, more than 77 percent of the world's mobile network data will be video and live streaming by 2021 with an estimated 1 TB of data an average mobile user will download in 2020 [3]. So, it is neither smart nor efficient to keep all our data transmission on a traditional wired network, which may enjoy more stability but less convenience.

Caching is not a novel idea, and the meaning of it has evolved in tune with our developing technologies. It was first defined as the process of storing copies of files in some temporary storage for users to gain quick access. Tons of different hardware have caches to cache, but the Internet has grown much more powerful, and this word, caching itself, is now often used for in-network caching. To make things worse, our established way of mobile cloud computing accessed by smart terminals has greatly increased the burden of network load and put higher demands on network bandwidth [4]. Methods like increasing base stations and acquiring new spectrum have been proven ineffective due to high cost with complexity and low scalability [5].

In order to resolve the limited computing along with storage capabilities and higher power consumption problems of mobile terminals (especially low-cost IoT terminals), it is imperative for us to shift high-complexity and high-energy computing tasks to the server side of the cloud computing data center [6].

Hope is that by doing so, the energy consumption of some low-cost terminals is brought down and therefore

standby time is extended [7]. But the shift of computing tasks to the cloud not only brings a tremendous amount of data transmission and computation, but also introduces a longer data transmission latency, bringing a fatal impact on some delay sensitive business applications (such as applications for industrial control and medical usage). These are what the 5G technology has promised, interconnectivity with lower latency and higher bandwidth. In addition to storing contents at several far-off top-level servers, mobile edge caching manages to store popular contents or fragments of contents on some edge servers near our end users.

Figure 1 shows us the general architecture of our edge caching networks. The core network serves as a bridge between us and the Internet these days. Base stations around us will make it more stable to communicate with servers more efficiently across our various devices. Cache plays a pivotal role in almost everywhere. Much as the authors mentioned in study [8], a redesigned framework will reduce the time of completing tasks, as well as latency, etc.

Besides from an overview of the network evolution from current 3G, 4G, 5G to 6G in the future, article [9] validated that deployment of mobile network caching possesses great potential in reducing redundant traffic. Besides being a positive candidate in alleviating the heavy burden of growing network traffic, edge caching brings lower delays and fair deployment costs as well, as we can see that mobile edge caching is sweeping across the globe as a core feature of the 5G network infrastructure. According to a report by Andrews et al. [10] that in 2020, the number of devices connected to the network will reach approximately 20.8 billion units, and the localized deployment of edge caching can effectively improve network response speed along with shortened network latency.

Harnessing the fact that mobile networks are now evolving from the conventional provider-centric structure to a client-centric structure, users will definitely experience an improved Quality of Experience (QoE) [11]. Speaking of which, Quality of Service is something similar but from a different perspective. QoS is defined as a measurement of the overall service quality that different kinds of properties come with different priorities. Five known parameters are packet loss, jitter, latency, bandwidth, and throughput. For QoE, though, it has historically emerged from QoS with more objective measurements with the experience of the end user being top priority whereas we did not have adequate capabilities of computing back in the days.

Consecutively, we make comparison and our contribution of this review in Tables 1 and 2, point out its novelty, explain its organization, and provide a table with acronyms frequently used in this paper.

1.1. Comparison and Our Contributions. Numerous works with novel insights of mobile computing and caching have been issued with the expectation to address various factors concerning edge computing and edge caching [26]. However, to the best of our knowledge, previous works done by other researchers fail to cover latest studies addressing QoS and QoE issues. Literature [13] took a closer look at mobile edge computing and caching, but they focused on application and synergies in between. The mechanism and algorithms of edge caching lacks deeper and thorough approaches with several updates missing in their work. Study [8] laid stress on three case simulations and lacked the overall perception of service quality. So, we think a comprehensive summary with detailed specifications would still be of benefit to others in this field. To fill this gap, we summarized the literature we encountered involving numerous properties of edge caching with QoS and QoE. Our work does not concentrate on the implementation of some techniques on ground level. We managed to take another approach by focusing on caching strategies with QoE metrics involved, which has other factors to be considered like the segmentation of popular contents and so forth.

Our main contributions of this paper are listed as follows.

(i) We provide a possibly comprehensive survey of mobile edge computing with caching combined. Studies we quoted are compared in different aspects with notable merits within

(ii) We manage to brief works related to edge caching and comparison between Quality of Service (QoS) and Quality of Experience (QoE), respectively, to further elaborate the significance of enhancing QoE in current network architecture

(iii) Caching process is divided into four stages just as four main focuses of QoE to fully analyze each corresponding disadvantage and offer potential countermeasures

(iv) Four novel applications and four crucial challenges are categorized via different focus of their requested services. Future research directions are presented correspondingly

1.2. Paper Organization. This survey is organized as shown in Figure 2. The rest of our paper is organized as follows. Section 2 is a briefing of works related to our topics, edge caching, and relation between QoS and QoE, with techniques used to better users' QoE following. Section 3 consists of an overview of current development with mobile edge computing and caching. Section 4 tells something about the four stages of caching and relevant issues with them. Section 5 comes with the QoE part, which is aimed at bestowing service providers with better network scheduling capability under numerous restrictions. Section 6 sheds some light on different user case scenarios and various applications. Section 7 is something about challenges ahead and future scope of mobile network. Section 8 closes this paper with our conclusion.

1.3. List of Acronyms. For better readability, we summarize all relevant abbreviations used in this paper in Table 3.

2. Mobile Edge Caching: State of the Art

Numerous studies have been carried out in order to achieve optimal performance of content caching, while QoE is still a

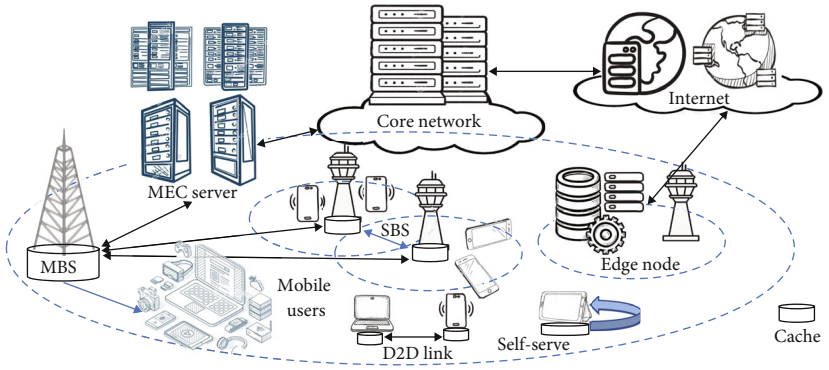

Figure 1: General architecture of mobile edge caching networks.

Table 1: Existing surveys on MEC topics and our new contributions.

Related works	Topic	Key contributions	Limitations
[12]	ML-based proactive caching	A comprehensive survey on machine learning-based mobile edge caching, with introduction of UAV-based content caching and applications of learning techniques	General content caching mechanisms lack thorough investigation.
[13]	Mobile edge network features	A survey of mobile edge networks, with detailed elaboration of computing, caching, and communication mechanisms as a whole, covering advances made possible with the synergy of computing and caching and key enabling techniques	This survey lacks recent advances both in caching schemes and development, with limited coverage on challenging issues.
[14]	Recent advances in MEC and content caching	A short survey of advances in MEC and content caching, including summary of MEC, cache replacement schemes, network optimizations, and open challenges	This survey lacks in-depth reviews of MEC and content caching mechanism.
[15]	Popularity-based video caching techniques	A survey on popularity-driven video caching techniques under cache-enabled networks, covering both single domain and cross domain features, with comparison of caching algorithms based on popularity prediction and different metadata models	This review focuses on one single application scenario and lacks coverage of other corresponding challenges.
[16]	Caching, analytics, and delivery of video caching	An inclusive survey on video caching, computing, and communications (edge-C3) deployed at the wireless edge, focusing on the fields of supporting components of video streaming and network technologies along with cutting-edge video delivery mechanisms. It also includes insights in the challenges of discussed context.	This survey focused on solutions of video streaming and delivery without the integration of MEC context.
[17]	Machine learning techniques for caching	A comprehensive survey of ML techniques application for caching in edge networks, including recent advances made in ML-based application on addressing the problems of prediction and clustering, thorough debate on ML-based edge caching, and summary of future challenges and research issues	This review concentrates on ML techniques in edge caching, but lacks full coverage of the originality of discussed issues.
[18]	Information-centric mobile caching	A survey focusing on information-centric mobile caching, including novel content caching policies, research development, and optimization analysis	Novel applications and recent advances of edge caching schemes have not been presented.
[19]	Content caching in vehicular edge	A survey on computation offloading and caching delivery in vehicular edge computing, including architecture, framework, and comparison of related studies	This survey focuses only on deployment of techniques of VEC.

Table 2: Existing surveys on MEC topics and our new contributions (continued).

Related works	Topic	Key contributions	Limitations
[20]	MEC in industrial Internet	A survey of key MEC technologies and typical applications in the industrial Internet, including requirements and related issues of deployment	This work focuses only on MEC deployment under the industrial environment.
[21]	Low latency caching solutions	A comprehensive review on low latency solutions of RAN, core network, and caching techniques, including waveform designs, multiple access, and transmission techniques in the field of RAN, SDN, NFV, and MEC/fog network architecture of the new core network and various latency reduction approaches surveyed, with appealing results of field tests presented	This review focuses on latency reduction approaches, which lacks general elaboration of other aspects in current network architecture.
[22]	Caching in content-oriented networks	A survey on research activities for caching networks, including technical issues like traffic engineering, congestion control, and security, along with design of caching network and performance evaluation	General elaboration and analysis of caching network deployment lack of deep exploration.
[23]	Performance evaluation of cloud service	A survey on cloud service evaluation from a systematic perspective, including reviews of both measure-based and analytical modeling-based approaches	General ideas of cloud service mechanisms along with architecture are missing.
[24]	Regulatory, standardization of 5G technologies	A survey of technologies incorporated in the core RAN and 5G network, aiming to increase system bandwidth, spectral efficiency, and RAN flexibility	Mobile network development with industrial demands lacks detailed elaboration.
[25]	Edge computing designed for IoT security	A survey on edge computing-based IoT security issues, including security solution, architecture, and privacy-preserving design, with outlines of several challenges in field	The paper only focuses on IoT security aspects in mobile edge computing.
Our work	Mobile edge caching, QoS, and QoE	An extensive survey on mobile edge caching schemes and developments, particularly (i) We manage to bring a comprehensive coverage on recent advances made both in mobile edge caching and QoS metrics, with detailed elaboration of MEC and edge caching architecture (ii) We extensively discussed the four stages of edge caching with top-down schemes of caching algorithms (iii) Insights into the integration of modern techniques and caching schemes are provided with light shed on research challenges and directions	Specific and concentrated reviews on key enabling technologies may not be covered and elaborated thoroughly.

new field embracing emerging techniques. QoS also needs to be included, as current literature neglects it when elaborating edge caching algorithms and schemes. To cover the basic concepts of our mobile network architecture, we will provide an in-depth summary on edge caching with QoE and QoS involved in this section.

2.1. Edge Caching. Caching techniques have already been looked at thoroughly, like web caching, content distribution networking (CDN), and information-centric networking (ICN). Back in the early 1990s, when the Internet was just another emerging technology, network congestion is a result of excessive data generated from various web pages and images, and that issue was resolved by web caching technique. This kind of caching helps store popular files momentarily on proxy servers or at users' PC. When it comes to the 21st century, videos became trendy and hence the congestion resulted from the transmission of videos. Problem was alleviated by the deployment of CDN. Subjects like the architecture design of caching, content deployment, and content delivery have already been studied in depth. Literature [27] elaborated the comparison between inefficient web caching which caused redundant data transfer and improved cache implementation. In the two datasets involved, redundant data contributes up to 20% of the total HTTP traffic volumes, with 9% energy consumption in total,

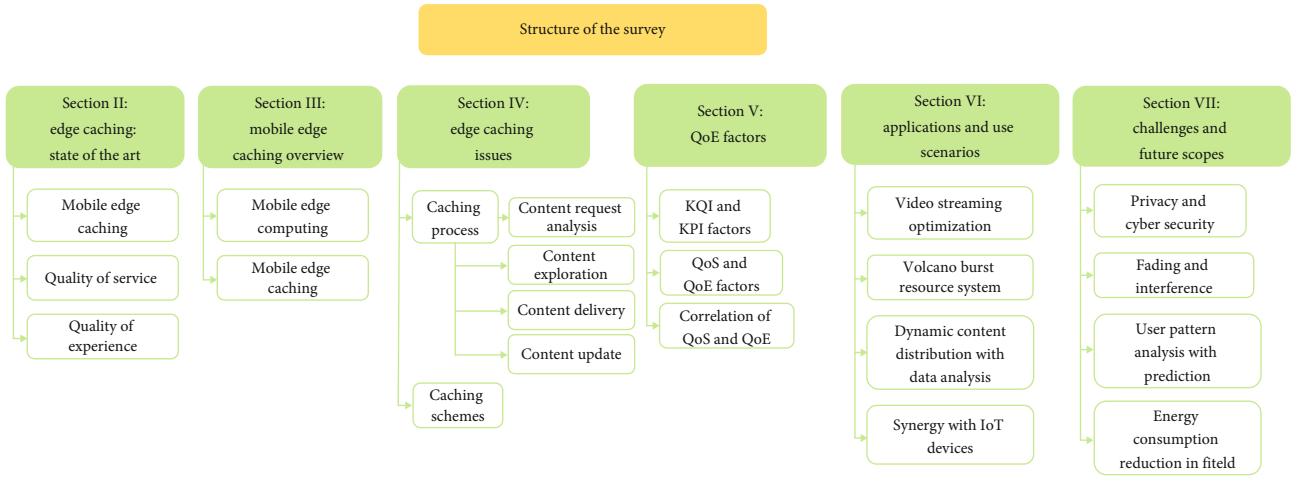

Figure 2: Organization of this review.

Table 3: Frequently used acronyms.

Acronym	Full terminology
5G	5th generation
BS	Base station
UE	User equipment
D2D	Device-to-device
IOT	Internet of Things
ICN	Information-centric network
CDN	Content delivery network
MCC	Mobile cloud computing
MEC	Mobile edge computing
QoE	Quality of Experience
QoS	Quality of Service
ISP	Internet service provider
SDN	Software-defined network
KQI	Key Quality Index
KPI	Key Performance Index
ML	Machine learning
RAN	Radio access network

which means caching strategies will make a difference in bettering efficiency and lowering energy waste.

2.1.1. Efficiency-Based Improvement. ICN focuses on identifiable information rather than a stable connectivity under traditional network architecture where hosts are in the center. In-network caching can help alleviate network traffic and is thus regarded as an appealing feature of ICN. Reference [28] brought us a comprehensive survey reviewing recently proposed mechanisms under ICN. Through extensive experiments and simulations, authors found out that reducing redundant contents is more effective and cost-friendly. ICN caching is still in its development phase, and many theories need to be validated as well. The authors in work [29] carefully studied the ICN architecture, with new features it brought with. Caching performance is subjected to many other properties, like cache size, caching policies, and fluctuating popularity of cached contents.

Mobile edge caching has improved systematic efficiency and can deal with tasks more efficiently than before, as we can see in study [30]. Caching action takes place at almost every location across the network architecture, from user equipment (UE) to base stations (BSs). Even edge replays are now empowered and can do some basic computation. Different caching schemes have susceptibility in almost every use case scenario. We also have different standards and criteria to meet in different cases. What is more, four processes of caching have shed some new light on our recent works and will serve as our main focus in our future research. Reference [31] first discussed some technical misconceptions of wireless caching. For instance, models which are aimed at uncovering popularity hit ratio should not focus on static ones because it fails to track variations and fluctuations. Others like wired networks and wireless ones should not be the same thing, or security issues are critical to the overall performance. Many aspects have been covered, and some promising research directions have been elaborated.

In pursuit of general MEC performance, literature [32] proposed a novel cooperation approach of computation and communication, with nearby helper nodes sharing resources actively. Based on joint optimization techniques, the authors brought one effective algorithm to map the optimal solution in partial offloading case. Numerical results elaborated the merit of the proposed cooperation scheme. Basic setup with single user and one helper can be extended and applied to multiple users and helpers.

Active and responsive mechanism does make a difference. Article [33] specified the notorious 4 V of big data, like variety, velocity, voracity, and volume. In their article, the authors addressed these issues with tradeoff between cost and gain in deploying proactive caching. Proactive caching discussed in [34] serves as one of the key enablers which are aimed at the future against backhaul congestion. To validate the effectiveness it promised, authors first proposed a caching mechanism with content popularity and correlations considered. Then,

they carried on showing reduced traffic demands with the aid of proactive networking and caching schemes. To be specific, work [35] presented PopCaching, which learns the popularity of the contents to be cached and makes decisions according to it. In their models, PopCaching smashes existing algorithms by 40% in cache hit rate, with low complexity for machines to learn. In [36], effective scheduling mechanism of resources in MEC server and BSs is discussed. They formulated the problem as a Stackelberg game. The game works as the MEC server collecting the tax while BSs compete for higher revenue and efficiency running on resources like cache size and computing power. The main purpose is to improve the QoE of end users. They proved that even a problem as difficult as a Stackelberg game can be solved with backward induction. Their proposed scheme has been validated by various simulation results. What is more, in [37], the authors also developed a game-based theory for video sharing APPs. Edge cache in this paper has against limited resources and differentiated QoE demands. Results are quite appealing with an optimized tradeoff like in [38].

Previous works always tend to separate MEC and D2D for different purposes, where MEC was claimed to resolve traffic offloading and D2D was deployed to engage in short range information transmission [39]. In [40], researchers came up with a novel idea of modeling behaviors of the requests and analyzing the popularity of contents with different techniques. Besides MDP and Zipf distribution, reinforcement learning- (RL-) based algorithms are also deployed to precache popular segments of files through learning, and some systematic adaptations were made to minimize the energy cost. Algorithms grow and evolve everyday like human beings, and their complexity grows exponentially. To conserve energy, we ought to reduce the complexity of algorithms while maintaining their performance in future works. The ideas of [41] are very much the same, but Zhang and Wang further updated D2D caching with centralized and decentralized combined. Their simulations validated that collaborative D2D caching schemes outperform others.

2.1.2. Popularity-Based Prediction. Content popularity has been addressed in studies, and there remain some problems. Caching subjects are chosen mainly according to content hit probity. In [42], they proposed a hybrid caching policy where popular contents and nonpopular contents are stored in different locations. Like the router finding the best path to the destination of its request, this probabilistic caching strategy has been proven to greatly improve overall system performance. There is also a sign of bigger backhaul capacity. Popular contents are reused in an anachronous way by many users according to work [43]. In this article, erroneous information may hamper the overall performance eventually, so some actions need to be taken to realize the full potential of edge caching. After determining the popular content, we also need to retrieve them. Deng et al. designed an algorithm in [44] to bring out the nearoptimal performance of allocating requests, with appealing experimental results.

Users' preferences are a result of big data mining and are usually kept from us before first appearance. How to cope with the dynamic nature of both users' preferences and network status becomes another priority. With the aid of online Bayesian learning, literature [45] showed us the true value of dynamic clustering policy. This policy has been validated with increased cache hit ratio along with faster convergence. Likewise, a self-learning cooperative edge caching scheme is proposed in [46], which focuses on different social characteristics among vehicles. To achieve maximized content dispatching efficiency, they will need to deploy diverse edge services.

To model content request and predict content popularity more efficiently, Mehrizi et al. in [47] came up with a model with great flexibility and adaptability. They captured the similarity between contents concerning network features via a multilevel probabilistic model. They also utilized Bayesian learning to obtain model properties with small numbers of requests. Similarly, in [48], Chen and his colleagues summarize recent works on AI and MEC combined. Afterwards, they deployed AI techniques in a case study and achieved appealing results. Their proposed edge service would prefetch 14 videos to obtain an overall rate of 90% hit ratio.

Mobility issues, which is entwined with content popularity and user preference, are the most complex and unpredictable among all of them all [49]. Even with the aid of modern computing techniques, predicting ones' behavior is not that easy. Take a city center for example, where so many people move in so many different directions every day. With such huge traffic, we are unable to make predictions and thus, we may end up with low prediction accuracy. This is not beneficial to the performance gain of edge caching and may incur some extra cost, such as the energy wasted in addressing backhaul traffic and deployments of more cells resulting from ineffective content placement. In [50], their team proposed a novel proactive strategy with mobility considered against ultradense network conditions. The probability of ongoing downloading is analyzed to provide information of maximize capacity effectively. Their simulation results are positive with a significantly lower delay of transmission and a higher hit ratio. Caching strategies must take every possible factor into account in order to work under any circumstance.

In [51], though, they proposed a comprehensive review of the prediction results from the prediction model as well as the predictability of cells. The common ground is that, for cells with low predictability, caching the most popular contents is our priority, whereas cells with higher prediction accuracy, our priority is to cache files abide by the results from our model. They suggested that we should implement different schemes when facing different network scales and complexity. Previous studies fail to consider other involved subjects and cannot be implemented to other complex situations where popularity of contents is unpredictable.

2.1.3. Resource Allocation Advances. Under real-life QoS constraints, study [52] came up with two low-complexity time allocation algorithms which enable energy-efficient

offloading and resource allocation under a multiuser MEC context. Simulation results validated the effectiveness of the WPT-MEC system, outperforming these systems without cooperation.

Then, to do something with various hardware limitations of users' devices, in [53] there is a concept of MEC meant to enhance device performance. And that means every available resource will be utilized without excessive infrastructure investments. Nevertheless, device-enhanced MEC will provide an improved and more steady QoE. Especially in dense networks, numerous end devices have their neighboring devices where D2D communication pervades. But as a nascent research area, there is still much work for us to take a closer look at and some might require thorough investigations.

Power consumption is another issue we must deal with. Article [54] first made an investigation of the changing nature of service behaviors in edge caching; then, a centralized model was formulated to maximize content caching efficiency while bringing down power consumption. Results show that Fang and his team can maintain service performance with a relatively lower network power consumption compared with existing solutions.

Storage is also of great significance when deploying various caching algorithms and schemes. In [38], Xie and Chen applied time-domain buffer sharing to improve storage efficiency. They scrutinized the tradeoff between storage cost and communication gain by comparing maximum caching time period. The algorithm tailored for this situation has converged with another one dealing with different users' demand preference. To achieve an ideal result, a two-layer searching algorithm was presented in their paper with simulation results validating its potential. In the best case, we can trust our system with intelligence for not only searching content items worthy of transmission back to the core network, but also echo times the transmission will take in the buffer of BS.

Intelligent connected vehicles are claimed to be the next gold mine waiting to be exploited. These vehicles might be of great help serving as relay nodes for data transmission. In [55], Deng and Xia first looked at the effects of outdated information on the Internet of Vehicles, and then they analyzed several parameters followed by some simulations. The results of these tests validated their theory of positive correlation between system performance and large number of cache relays. For future works, they will consider new materials and new techniques to enhance performance in networks. Reference [56] brought us not only a concise review of D2D-assisted caching performance, but also a detailed comparison of tradeoff between different schemes. And the conclusion of the paper is appealing that the D2D network has competitive performance with cheaper untapped and developing resources compared with our current one under service.

2.2. Quality of Service with Quality of Experience. Literally speaking, experience is more from a user's perspective with emotions and prejudices. Defined by ITU-T SG12 (2007), QoE means "the overall acceptability of an application or service, as perceived by the end user," while service's quality is more solid, which does not care whether the users feel content or irritated. And there has been some literature addressing the correlation between QoS and QoE. We must point out that QoS is only a part of the picture, and there are more properties emerging to define the whole network system with a surge of studies addressing various problems with the hope of improving QoE from users' perspective. Several factors are impacting the overall QoE and themselves as well. Studies have exploited some parts of these factors, and we managed to categorize them as follows. Table 4 shows us the QoE influencing factors, and to be specific, technical factors are these influencing QoS. Nontechnical parts are more subjective and are varying from time to time. These subjective factors are now the crux of the whole network system that requires more attention.

In [57], the authors compared and categorized some of these factors shown above. QoS was divided into four parts as we discussed before. Routing, traffic shaping, resource scheduling, and traffic control each plays a vital part in the whole network system. Quality of Service was divided into objective quantification and subjective quantification. QoS is more fixed on the network, the process of transmission, while QoE attached greater importance to the end terminal, our users. One more difference is whether the measurement is subjective or objective. In [58], the authors held the view that QoS and QoE are interdependent to each other. And the comparison between two groups validated the point. QoE and user's perception change simultaneously with parameters of QoS fluctuating. The authors also pointed out the future direction, like focusing one key variable at a time.

2.2.1. Content Delivery Improvement. Video is the dominant king in mobile traffic with live streaming and vlogs growing in popularity across the globe. Video-based services and their considerable revenues have forced the service providers to upgrade their services with higher quality. The authors from literature [59] did a great job by reviewing the evolution of quality assessment methods. Subject tests and objective tests each have their own flaws and are hungry for an update. Thanks to our big data analysis, new models with new metrics can be applied to a larger-scale system. Current infrastructure like MEC storage and edge BSs comes at a great cost and cannot provide enough resources and optimal QoE because of ineffective resource management. The whole system needs an update. In [60], Kim and his team managed to improve the cache hit ratio about 12% with backhaul bytes ratio down about 12%. Average delay was also hopefully reduced by 14%. The STV request model with DASH streaming combined has proven capable of satisfying users' QoE and bringing down backhaul traffic.

In a multimedia multiservice scenario, as considered in literature [61], QoS fluctuations are investigated through diverse characteristics, along with a QoS evaluation model designed to maximize QoS. Authors applied Deep-Q-Network technique to achieve flexible resource allocation, followed by validation of QoS-aware resource allocation effectiveness. Better performance in resource allocation was achieved compared to other algorithms.

Table 4: Technical and nontechnical factors affecting Quality of Experience.

Technical factors influencing QoE	Nontechnical factors influencing QoE
QoS guarantee mechanism (end to end)	User behavior and QoS of ISP
Key Quality Indicators (KQIs) of operational performance (end to end)	Convenience and easy access to Internet providing services
User's ability to reach and transmit	Content of provided service
Network/service coverage	Price and customer support
Functions and performance of terminals	User tolerance and behavior

With the aim of handling real-time (RT) traffic more efficiently, authors in study [62] proposed an approach to classify user traffic into four different classes: conversation and streaming in RT traffic, along with interactive and background communication in NRT traffic. With MEC servers buffering each traffic class individually, RT traffic is prioritized with more resources than NRT, which in turn improves general QoS metrics, with reduced jitters, delays, and higher throughputs.

Literature [63] employs an emerging model of Cognitive Agent (CA) and utilizes LSTM to make prediction of terminal behaviors and requests, in order to better collaborate resources among devices. Performance was evaluated through hit rate and task execution delay, which showed better QoS and QoE, respectively.

2.2.2. Caching Mechanism Improvement. Authors in literature [64] reshaped current recommendation systems, utilized them to nudge content demands, and optimized user-centric experience and network-side performance. To achieve these objectives, they formulated a joint optimization problem of decision-making and achieved possibly highest cache hit ratio with an enhanced QoE for end users.

In article [65], the authors managed to integrate SDN (software-defined network) with a new vehicle edge computing framework, providing different QoS accordingly. The interaction between vehicle and the edge server is designed as an incentive mechanism based on Stackelberg game modeling, with a corresponding optimization algorithm proposed. Experimental results have validated the effectives of their proposed scheme, with a promising 76% reduction of delay compared with conventional edge computing scheme.

A three-tier architecture with users, edge servers, and remote cloud servers is proposed in study [66], with the capability to adjust caching decisions in line with popular patterns of time and space services. System delay and lower energy consumption are achieved in simulations.

Authors in literature [67] deployed three hierarchical caching schemes to meet different motivations. The random hierarchical caching mechanism is aimed at maximizing average caching capacity, and the proactive caching mechanism achieves overall higher cache hitting rate with the game-theory-based caching mechanism minimizing general caching expenses and aggregating available capacities of all three caching ties.

In the case of cost-service caching without resource sharing, literature [68] proposed a randomized rounding algorithm with the analysis of approximation ratio. Game-theory mechanism to address cost and delay-sensitive service caching was also proposed and validated the idea that cost reduction is achievable through collaboration with others.

2.3. Lessons Learned. In summary, recent years have witnessed numerous bright ideas on improving QoS with much assiduous work done. There are many fruitful results, and the literature mentioned above are just the tip of the iceberg. As can be seen from Figure 3, we are about to experience a great leap in edge caching with QoS techniques. The reason for QoS being off track is for the concept of decentralization. Machines do their jobs these days while people are at the end of the line in this whole operation. We used to centralize powers to make our computation tasks easier to resolve, but new techniques have changed that obsolete notion with everything being mobile nowadays, with scattered computation capabilities.

3. Overview of Mobile Edge Caching

When it comes to mobile edge caching, we can hardly ignore the fact that demands coming from our users keep rising, thus making our traffic overloaded. Therefore, it is an arduous task to schedule and manipulate these overloaded networks. What is more, many emerging applications need to feed on services with higher reliability and stability (i.e., higher data rate, lower latency goes with better QoE). Clearly, we expect our current network architecture to keep in line with the time and thrive in the future. After all, many protocols were put forward several decades ago. In this section, we will elaborate how MEC (mobile edge computing) facilitates the development of mobile edge caching. You will also see the review of the benefits of MEC and mobile edge caching to make way for improvements of each related factor.

3.1. Mobile Edge Computing. To cut latency and response time of provided service, mobile edge computing lifted our experience by computing in a closer range of our end devices. Literally speaking, edge computing involves algorithms that utilize the margin area which is in proximity to the clients. Unlike cloud computing which employs centralized big data processing, edge computing equals decentralization. Cloud computing does have many advantages, such as powerful computing capability with huge storage capacity. Yet people cannot endure the long latency it brought with any longer [69]. Figure 4 shows us the role fog was supposed to play in our system, but edge nodes are taking its place. Fog computing is now like the bridge, a smooth transition

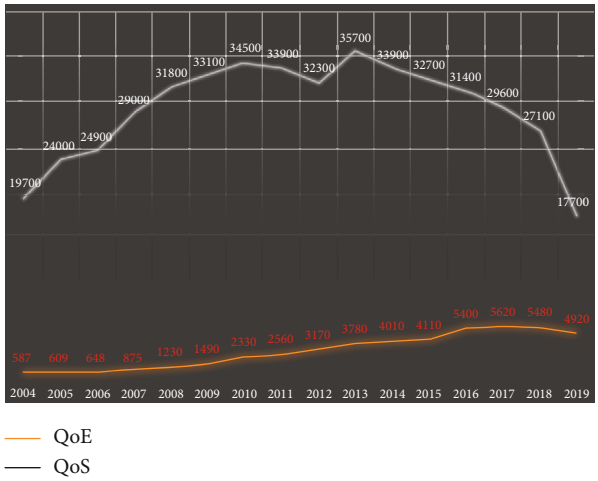

Figure 3: Comparison between numbers of papers with QoS and QoE focused provided by Google Scholar.

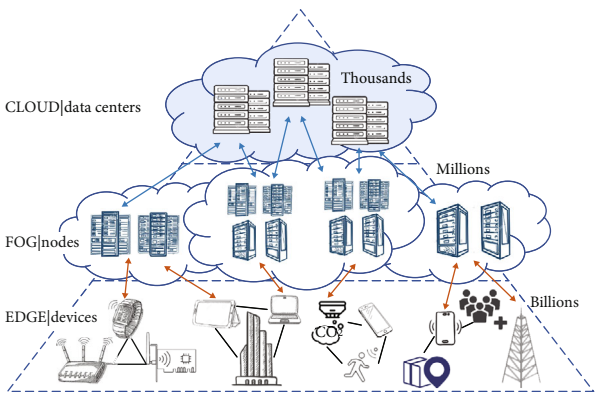

Figure 4: Cloud-fog-edge architecture manifestation.

that connects cloud and edge, and fog solutions usually come with an extra price of infrastructure deployment [70]. So here comes the MEC.

MEC takes one step further, by possessing several unique advantages against cloud computing. First, thanks to the focus on real-time, short-cycle data analysis, MEC enjoys lower latency. Second, adjacent to the end user means higher efficiency while filtering and analyzing data at the edge nodes [71]. And a lower data response time is achieved through data processing via edge nodes by cloud transmission. Third, since more and more computing tasks have been offloaded to the cloud server, we can cut the cost down to 60% compared to cloud computing alone [72]. Since devices are free from depleting their precious computing resources, it runs on lower energy, according to Kiani and Ansari [73]. Fourth, better services are available for end users since they are closer to our caching edge. Our service providers can now come up with numerous personalized services to offer [74].

European 50 PPP (50 Infrastructure Public Private Partnership) has recognized these unparalleled advantages of MEC as one of the most principal technologies designed for 5G [30]. The whole system of MEC can merge in the existing network system reconfigured for future 5G (which is specified in work [4]) simply by upgrading network functions via software means.

3.2. Mobile Edge Caching. We see the Internet growing at a speed far outpacing our relevant techniques; something must be done to curb its aggressive invasion. Caching techniques which store contents (such as popular videos and inviting webpages) at several storage locations not far from the mobile network edges (base stations or user equipment) for future usage have the potential to reduce backhaul traffic and better QoE of users [75].

Mobile edge caching is envisioned as one of the most promising technologies owing to its efficiency to reduce traffic congestion. According to Wikipedia, caching is defined as the process of storing data in a defined cache. Edge caching is a significant use case of MEC where future requests for that data can be transmitted to requested users smoother and faster. According to a study published in 2018 [76], mobile edge caching is likely to reduce backhaul capacity requirement by at least one-third via working out the similarity between requested contents or segments of popular contents. In our traditional network architecture, mobile users send requests for specific contents to a centralized Internet content server that was usually heavily occupied and even too far away from us.

When retrieving the same popular content from remote servers, users' devices have to make repeated requests which can lead to traffic congestion. Well, with over 3.3 billion mobile users in 2018, we see mobile edge caching as our next big thing. In a typical mobile edge caching model, content requests issued by users' devices are first received by one of the numerous edge nodes around the user. And the domain name system or namely DNS reroutes the request to the nearest and possibly fastest cache. Problem was solved. Moreover, thanks to the lowering cost of storage, it is a much more cost-effective way to deploy cache at the edge.

There are several merits coming along with mobile edge caching. First, it enables the reduction of latency when retrieving the users' requested contents. Second, it helps to reduce backhaul links, with significantly lower backhaul traffic. Third, as a promising use case of MEC, mobile edge caching possesses some of MEC's advantages, such as lower energy consumption in consort with reduction of cost [77]. Fourth, mobile edge computing can also make use of the data collected by the edge servers, and by analyzing it, improved computing efficiency is expected.

3.3. Lessons Learned. What we can learn from above are plenty of reasons for us to embrace MEC and edge caching. Fog computing was also trendy around 2017 and 2018, but it has some fatal drawbacks and fails to live up to our expectations. Likewise, MEC and edge caching also have their own flaws and need to be fixed with different mechanisms. There is a more specific summary focusing on MEC and its collaboration with many applications in literature [13], while our work puts emphasis on edge caching instead. Mobile edge caching is only one piece of the puzzle but also has a great impact on QoE, and we will then take a closer look at it.

4. Issues concerning Edge Caching

Edge caching is deemed as a mainstay in future communication network systems across the industry. To fully realize the potential of mobile edge caching, the following key issues should be studied thoroughly: what content to cache involves the popularity of requested content and awareness of its patterns, where to cache refers to the very location that caching takes place, and how to cache has something to do with deploying caching schemes. And cache performance deals with caching mechanisms and algorithms [78]. The performance of mobile devices is often impoverished due to the limitation of computation, storage capacity, poor battery life, and other constraints.

To better address these restrictions, one creative solution we mentioned before is offloading some computation tasks to the cloud [79]. Storing the requested contents in advance will help shorten the response delay of a user's specific request, which means lower latency and in turn brings better QoE and alleviated backhaul pressure [80]. But researchers have taken different approaches to handle it. In this section, we will introduce some advances made by researchers and investigate the mechanisms that provide caching at the mobile edge with more productivity.

Table 5 summarized related studies focusing on different working areas of caching schemes, and there has been more work focusing on cooperative policies with novel applications of machine learning and deep learning techniques as discussed in Section 2.

4.1. The Caching Process. The caching techniques in cellular networks are merely in its infancy while we have done enough academic research and algorithms design. We all know that edge caching has two expectations; one is to improve users' QoE from their perspective, and another is to make resource scheduling and management more efficient to alleviate traffic congestion from the network's perspective [6]. And saving some energy and cost would also be nice. Unfortunately for us, it is quite difficult to grab all the advantages above, and that is our ultimate goal [81].

We will then open doors for the discussion of the caching process, including the four main stages and other key issues related to the actual performance.

4.1.1. Content Request Analysis. The first phase of caching is called content request analysis. Technically, contents to be cached are mainly files and videos. Files are most likely to be time-tolerant while there are some exceptions. However, when it comes to videos, online streaming and real-time videos are usually time-sensitive. What is more, the IoT data tend to have a shorter lifetime but their devices tend to run on lower batteries for recurring activation. Of course, merely IoT data alone cannot be convincing enough to manifest the real ambient environment standing. Hence, the freshness of the IoT data should also be included when deploying caching strategies.

When it comes to assessing the content placement problem, the content popularity becomes a requisite factor which reflects the probabilities of users requesting certain types of contents directly [82]. To categorize the content popularity more effectively, Zipf distribution is widely adopted [83]. Yet Zipf may not be as accurate as it was supposed to be when facing different types of files to cache, such as IoT data other than videos [84, 85]. Machine learning and deep learning could be applied to make a difference to measure the content popularity more efficiently; they will be helpful in some regression and classification models. Data mining and big data analysis may be employed to train the algorithms to be a better observant of users' history requests [86].

Popular contents grow and change nowadays in a second, and their properties and parameters change accordingly, the content popularity and user mobility likewise. Caching techniques have to adjust and adapt to the fluctuating nature of contents [87]. Coding and multicast are of great importance for obtaining useful information from caching. There is a case where requests and users are in the overlapping coverage of several BSs. What if every request is sent to one of the most preferred destinations? That will bring about a significantly deteriorated QoE and huge waste of computing resources [88]. However, for a small number of users in the network coverage, content placement methods may not be as effective as we can see among a larger group. That is because content popularity is a probability and does not necessarily work on one specific group of people [89]. We have to focus on other characteristics like the history of user's preference and mobility patterns.

User mobility pattern has huge impacts on the strategies of content placement as well [90]. Mobile networks develop at an unprecedented speed; the latency we aim to eradicate is a problem created by us when designing different or even complex topology of the network. All the complexity, inconvenience is attributed to the constantly changing nature of users' mobility [91]. Most studies see the Markov chain as the characteristic of the user mobility patterns for that the position in the near future is correlated and mostly influenced by current position [87]. Our patterns show our interests, hobbies, personality, and even social connection. So, users with more mobility patterns alike tend to share similar social relationships.

Hopefully, in [92], two effective self-adapting algorithms were proposed. Two constraints like traffic congestion and radio frequency have to be met in response to users' demands.

4.1.2. Content Exploration. To determine whether the requested content has been cached at one of the nearest edge storage, we must search the entire network to find out. Like routing in a conventional network architecture, we must design a technique to search for contents stored at every possible cache storage and find the l way to cache at little cost [93]. That is defined as the content query problem. Like routing techniques, it first sends queries to neighboring UEs, then searches for BSs, and finally walks across the entire coverage of the remote network. Then, in the worst case, these requested but uncached contents will be fetched from central servers through links with finite capacity. There

Table 5: Summary of literature on caching policies and algorithms.

Work area	Related work	Key points
Conventional policies (noncooperative included)	[100–103]	(i) Least frequently used (LFU) (ii) Least recently used (LRU) (iii) Most popular video (MPV) (iv) Distributed caching leaving caches in other cells untouched (v) Cache replacement problem with some selected models
User mobility-based policies	[100, 104]	(i) Local content popularity considered (ii) User's preference towards specific video categories
Learning-based policies	[103, 105]	(i) Estimation of timely content popularity with reinforcement learning (ii) Q-learning-based cache replacement strategy (iii) Coded caching scheme
Cooperative policies	[13, 90, 106–111]	(i) Cache management with cooperation among BSs and UEs (ii) Minimizing resource cost while maximizing data traffic from caches (iii) Collaborative video caching and scheduling among cells (iv) Tradeoff between caching redundancy and diversity (v) Joint caching and routing design

is a slight difference from routing in that the caching scheme favors these caches that have served requests beforehand; in another word, it prefers acquaintances.

4.1.3. Content Delivery. Moving forward, requested contents are delivered to end users via edge nodes or from upper level service providers. In the third phase of caching, we deal with the transmission back home [94]. The terminal stop of our transmission and band frequency are among the most principal issues [95]. Parameters including consumption of resources like power, channel allocation, and other required transmission properties of the edge caching system should also be elaborated.

4.1.4. Content Update. Some outdated contents with nearly zero popularity are meant to be discarded to make room for new contents to fill in. If a system dwells on the past, we are about to experience downgraded experience from the user and less efficient system, then. Hence, timely content updates at certain intervals are of great significance.

To summarize points above, it is neither possible nor necessary to cache all contents in servers both at the edge or in the center of the mobile network. There are so many issues involved, and negligence of one small interruption may deteriorate the whole QoE. More work should be done addressing the synergy of all small issues and conducted more compatibly with our reality.

4.2. The Caching Schemes. Caching techniques storing popular contents at edge storage nodes which are in close proximity to the mobile network edge have been proven to possess the potential to alleviate backhaul traffic significantly [96]. To our knowledge, the transmission of duplicated popular contents to various users leads to backhaul traffic in huge volume [97]. Contents are created with default popularity, but when they are uploaded on the Internet, most of the popularity begin to fluctuate. Popular contents attract a great deal of attention and are requested by a large number of different end users in different time periods. So, the network plays a dominant role by sending the same content to requested devices. Hence, the backhaul links bear too much duplicated data and that results in a significant backhaul traffic [42]. However, when we adopt the concept of "decentralization" and start to cache popular contents in the base station effectively, there is a small chance that some memory is left to store the precached popular data. Then, things will be a lot easier for backhaul links to send something popular to the BSs once, which can also be helpful in reducing a great amount of duplicated data and improving efficiency of both Internet and the energy consumed.

Some temporal but regular characteristics should also be taken into account like different time periods in our daily life as different user case applications, to refine the user behavior prediction. What is more, a system model for joint prediction of movements and behaviors combined should be proposed to make it more felxible in our daily deployment.

5. Factors Influencing QoE

QoE (Quality of Experience) can be understood as user's experience and perception, that is, measuring the quality of the network from the perspective of the user. When referring to QoE, we can hardly neglect QoS, which can be considered as a more objective version of QoE. ITU-T Rec E.800 originally defined QoS as "the comprehensive effect of service performance that determines user satisfaction [98, 99]." Specifically, QoS is narrowly understood as the KPI (Key Performance Index) of the underlying packet data transmission. These parameters include network delay, jitter, bandwidth, and bit errors. From the perspective of network equipment, it is easy to monitor and statisticize, and KPI reflects the performance of the equipment objectively [57].

However, only with these key parameters of KPI are not enough to reflect current network quality. The main idea of using it is to measure the user's feelings from the network's perspective. When using the KPI indicator system to

Table 6: Four key aspects of the whole network architecture.

QoE	Improve application delivery quality and bandwidth utilization value	Traffic limitation, bandwidth reservation
QoS	Reduce customer complaints and increase customer satisfaction	Quality perception and customer service support
KQI	Optimize network applications and improve business quality	Traffic sharing and application diversion
KPI	Reduce network risk and improve network reliability	Bypass monitoring, link backup

measure network quality, it is often the case that the KPI indicators of the entire network equipment are perfect while complaints from users are spiking. In order to further improve service quality of the network, KQI (Key Quality Index) is introduced to evaluate system of network service quality, which mainly focuses on service quality parameters that are more relevant to the user experience regarding different services. The essence of KQI is the end-to-end service quality of some key services, such as the smoothness of the playback, sharpness of the video, and synchronization of audio and video in video delivery.

KQI has improved the quality of key business services to a certain extent, but customers are also becoming more demanding. Factors affecting user satisfaction include a variety of content. In addition to the quality of the network, the speed of terminal opening pages, video continuity, personal sentiments, and feelings of fatigue also affects customer's satisfaction. We can see from the Table 6 that each of the proposed aspects has its own concern but together, they reflect the general experience of current network.

In this section, we will categorize some of the factors influencing overall QoE with the four metrics we mentioned above.

5.1. KQI and KPI Factors. KPI, as a direct agent influencing KQI, is a complex and opaque mathematical combination of various parameters named as performance counters.

5.1.1. Caching Subjects. Concerning the subjects we are caching, common content types include multimedia files (mostly videos and files) and data from IoT, which tends to have shorter lifetime and more dimensions [78]. Un-1045 popular contents, however, have a negative impact on the utilization of caching and may introduce duplicated traffic while most of the end users request the same popular content. Without the knowledge of the popularity of the requested content, no decision can be made [112]. What is more, in order to maximize the hit probability, user preference should also be taken into consideration to make optimal decisions. The hit probability of cache refers to the ratio of the number of cached files requested by the end users divided by the total number of files stored in the caches. And we may be able to see the increased hit rate by incorporating more parameters and dimensions into the current model. Instead of applying traditional caching schemes, it is a must to analyze the reputation of requested content, such as the concepts of the least recently used (LRU), least frequently used (LFU), and first in first out (FIFO) in our conventional network system [113].

Users' own behaviors like moving patterns in the parameter with their taste change from time to time, and that makes the current caching techniques not as effective as claimed or even obsolete. However, the patterns of content requests we gain by analyzing statistics should take both the mass and the individual into account while addressing the content placement problem [100].

5.1.2. Content Popularity. The contents with popularity above average in the network are not fixed and their patterns are always kept from us. They are also highly correlated to the user behavior pattern but has its own traits. But we would like to point out that the preferences for popularity and prediction results are to be adaptive and flexible. And future models with optimization should be built based on it [114].

5.2. QoS and QoE-Related Factors

5.2.1. Various Device Limitations. The conventional network system needs upgrades and evolution, so do our own devices, which also have some flaws and need to be dealt with [115]. We pick up our cell phones quite frequently nowadays and hardware limitations such as poor battery life have forced us to depend on our software by optimizing our resource allocation. There is a solution to cut the cost. Our devices will not need to wait for data coming from a long way from central data centers; base stations (BSs) will serve as a middleman. Caching at BSs will greatly alleviate the backhaul traffic and thus reduce the consumption of resources like energy [116].

BSs have the potential to compute and analyze, so we can take advantage of that and let the BSs make some predictions of popular contents beforehand on their own. After that, they might be able to serve end users without the need of backhaul links and thus improving the QoS [117]. This is of great significance during peak hours. Another thing about BS caching is that the coverage is sometimes far too small to provide services to potential customers [118]. Adding the interference and fluctuating status of wireless connection, we have to think otherwise like implementing more BSs in other forms, like MBSs, SBSs, PBSs, and FBSs [119].

In short, devices we are holding or some serving us from far away have a significant impact on our end experience, and they are also desperate for an upgrade.

5.2.2. User Behavior Including Mobility Issues. Due to the smaller caching capacity mobile end users are bestowed with, accumulated with other annoying issues such as the mobility problem and different topology of networks with inferences from others, we are gradually becoming user-centric [120]. We have hardly acquired the knowledge of human behavior prediction, let alone applying it to our algorithms. But here in [121], users' heterogeneous preferences

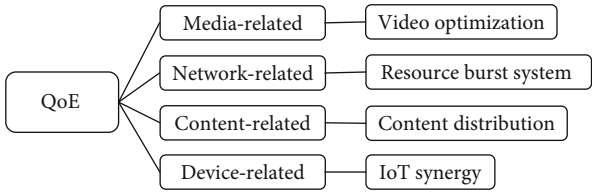

Figure 5: Relationship of QoE applications.

make it more difficult to cache some useful contents beforehand. After some bad experience of the network which keeps them waiting, some users may change to another channel or even abort all their requests immediately.

5.3. Correlation of QoS and QoE. KPI, as a direct agent influencing KQI, is a complex and opaque mathematical combination of various elements named as performance counters. In work [122], the authors first studied the relationship between KQI and KPI in video streaming services with VoLTE afterwards. They furthered their work with focus on the mathematical interdependence between QoS and QoE. Equation (1) can be expressed as follows with KPIs being QoS_i:

$$\mathrm{QoE} = f(\mathrm{QoS}_1, \mathrm{QoS}_2, \cdots, \mathrm{QoS}_n). \tag{1}$$

We can see from here that the QoE output has numerous KPI factors involved, considering various use case scenarios. And later, the authors specified that for video streaming service only, works can be summarized into Equation (2).

$$\mathrm{QoE} = e^{(\alpha_0)} e^{\alpha_1 \mathrm{QoS}_1 + \alpha_2 \mathrm{QoS}_2}. \tag{2}$$

In (2), the linear regression factors are represented as i, and as validated in previous work [123], QoS_1 refers to IP Packet Transfer Delay (IPTD), whereas QoS_2 refers to IP Packet Loss Ratio (IPLR).

In summary, the variants of user requirements keep changing and our future work will address them accordingly. In different use case scenarios, we will come up with relevant models and schemes against them as we do in edge caching.

6. Applications and Use Scenarios

Four different pieces of the puzzle have been mentioned in Section 5, and each of the novel applications we mentioned here focuses on one aspect of the whole network experience. Figure 5 depicts the key points of applications and relationships in between.

6.1. Video Streaming Optimization. As we can see, vlog and live video streaming are pervading people's lives around the globe, with vloggers becoming another lucrative occupation. Their demands on higher resolution with frame rates and smoother experience have taken up most of the traffic on the Internet. Video transmission accounts for over three-thirds of the total mobile traffic, and that percent has a steady rising trend [124].

Literature [125, 126] give us some new insights into online video caching with limited future information and various constraints. Both works incorporated online algorithms for optimal decision-making. Wang et al. classified the video popularity problem into a many to many matching problems with functions like dynamic adaptation of video bitrate, while in [127], caching techniques yielded different parameters of video contents but with chunking and dimension as priority. And with elaboration of the great effect chunking can bring, layering became the second candidate to greatly reduce energy consumption against some minor defect in performance.

The deployment of edge caching mechanism avoids wasting network resources on redundant video streams [36]. The deployment of MEC servers at the edge node has been proven to possess more capability than these at the video source. Caching at the edge is expected to enable dynamic content delivery with gathered information of network status and users' condition [128]. Since the content is placed in closer proximity to end users, we expect a more user-friendly QoE compared to the conventional architecture. And in the second example of study [8], authors proposed and utilized collaborative processing strategy, resulting in reduction of backhaul traffic load with lower occupied cache capacity.

Figure 6 below lies a conventional video delivering system on edge servers. Service providers rent edge cloud and cache videos on edge servers, while users obtain videos from nearby edge clouds when edge cloud precached the requested videos. When there are few requested videos or higher resolution videos cached on the edge server, users will choose to request them directly from the service provider.

6.2. Volcano Burst Resource System. People nowadays have powerful devices to play their favorite movies in 1080 p (FHD) or even 2160 p (QHD), and China has its own vision of deploying 8 K VR live streaming during the Beijing Winter Olympic Games. A higher resolution makes people feel content but it makes our network system burdened with gigantic amounts of data. With analytic data provided by Tencent, we can see that people now expect online video to be played not only in 1080 p but also with 60 fps on average.

And when it comes to entertainment, people may not be satisfied with movies alone, and they want to play games with high-resolution screens to get an immersed experience. That means we must be well prepared when the night falls with peak hour coming, and we are talking about scheduling 10 Tbps bandwidth. Of course, the peak of these large businesses will not be long, but we should get prepared. With the volcano system predicting current available bandwidth, we can have the capability to schedule resources up to 1 Tbps against emergency. What is more, with the aid of deep involvement of MEC, users can be used to bring down 40% of the network delay without further refinement [129].

6.3. Dynamic Content Distribution with Data Analysis. Our 5G's proposition includes bandwidth higher than before, but can we sense the difference when writing an E-mail or

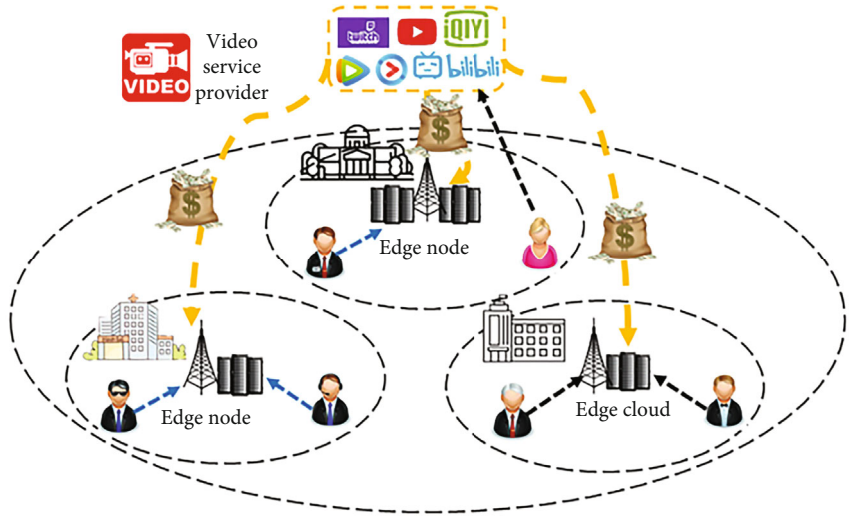

Figure 6: Illustration of video caching and processing with the aid of edge node.

just checking out New York Times for a piece of news? The answer is negative. Daily routines with minimum expenditure do not require a higher transmission speed or less latency. Some of them even share the same steps like loading a sign-in interface or verification of identity. But we have already paid the price, and our service provider will not change the speed to meet our requirements for us automatically. Applications with claimed cognitive assistance are deployed to augment our perception. Once we have closer and more stable access to our edge cloud server, we can make some of our latency sensitive applications possible by bringing down latency as much as possible. When the end user steps out the coverage of his original BS, he will be recognized by the system and handed to another BS with the optimal service. And the algorithm will not bother to look for another one for when the user has his history pattern analyzed [130].

6.4. Synergy with IoT Devices. Industrial applications have been craving for the universal usage of an upgraded version of network for a long time. Smart sensors make things even closer to our vision. Fog computing plays its part well in supporting the data transmission of IoT devices [131]. Sensors are merely in its development phase and cannot handle or store large volumes of data. So, we only need to take care of the rest, velocity and variety. Velocity involves issues of transmission as we discussed before. Variety needs to be recognized before data processing, and the characterization of big data is never easy. After caching techniques were introduced to IoT networks, works have proven that caching brings about significant reduction of energy consumption and bandwidth usage [132]. System delay can also be minimized after an energy efficient caching strategy is designed and presented in [133]. Mobile edge caching will keep inspiring the whole industry with various novel thinking springing and enabling new services provided by IoT devices.

One of the visions is to have every IoT device connected with each other in future network. With light-weighted wireless terminals, we can have stable access to the cloud and edge services and enjoy these smart real-time services without worries of depleted computation capabilities [134]. And for obvious reasons, data centers offloaded on the cloud have higher latency and fail to live up to our expectation. With the aid of edge computing, we can resolve this smoothly [135].

6.5. Lessons Learned. Optimization of video streaming is meant for enhancing the QoS and improving efficiency of resource scheduling, by the means of setting limits on traffic and reserving bandwidth in advance. The volcano burst resource system is more capable of handling a sudden surge of huge traffic and will lower the risk of cashing the entire network thus bettering our KPI with a higher stability. Dynamic content distribution is aimed at improving KQI via traffic sharing and application dispatching. The QoE part has something to do with customer service support. Synergy with IoT devices is more of the icing on the cake which enables us to take one step closer to our future.

7. Challenges and Future Scopes

Above are recent development made possible by edge computing and caching. However, there are several ailments need to be looked at. We cannot take network resources as the same thing as we used to presume in the old conventional way which is specified in work [136]. Parameters including power allocation channel allocation and other required transmission property should also be taken care of. In this section, we will discuss some of the challenges we may face and shed some light on future research directions.

First, various mobile users tend to have different or even contrary social interests. And that makes it impossible for one single caching scheme to work on everyone. Even for one person, he or she may have different interests during different time periods. So, it is imperative for our algorithms to possess some kind of intelligence to know the exact time

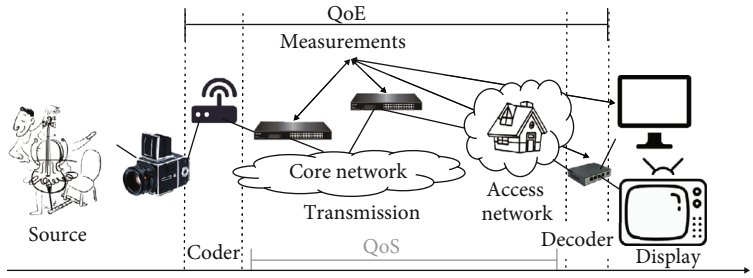

Figure 7: Illustration of QoS-QoE covering domains.

spot to switch to other more compatible policies. Second, we all know the super low latency our 5G network has promised us, but different clients have different demands for latency or speed. They may not expect to feed on higher speed since current 4G services are already teemed with data and news more than their mind can digest. We are expected to lower the price of higher speed and bandwidth, but right now, we must separate clients via their demands. Besides that, users' mobility awareness is another aspect that sometimes has a deadly impact on QoE. Users in this sector of the network are likely to be travelling around, and their requests are expected to be handed off to another cell frequently.

Figure 7 above illustrates the measurement of QoS and QoE. Data travel a long way from the source to the display for us to enjoy, with QoE being much more considerate than conventional QoS. The interconnectivity of 5G will also provide us a seamless and always-connected service.

7.1. Privacy and Cyber Security. Security issues have been a long-standing aliment among our system because wherever we go, they follow. Eradicating them is impossible but we can do something to calm the situation. Cyber security issues result from the abuse of the information of UEs, and that is almost inevitable after numerous data analyses [137]. Our conventional security solutions towards them fail to keep up with the pace and even backfire causing catastrophe. Security protocols now in service are based on full connectivity, which is harmful for mobile networks as too many links are intermittent as shown in literature [138]. So, we will have to redesign security protocols with some tailored improvements hopefully.

Privacy issues are catching its populace around the globe. User's personal information is vulnerable against repeated D2D communication with each other [139, 140]. Our devices want to hide personal information or make them desensitized. Tech giants like Google and Facebook are abusing it without us notified, while pervading exploitation of mobile data free from invasion of man's privacy remains a challenge for digital tycoons [141]. Tradeoff between convenience and privacy is always an ongoing topic among all fields [142].

7.2. Fading and Interference. To the best of our knowledge, many published works neglect the impact brought by channel fading and other interference [143]. The high speed we enjoy nowadays makes network topology more dynamic and complex. That will possibly lead to a less stable channel with much more interference around. We may detect some false signal of rising cache hit probability for loss of energy efficiency. That is because in reality, the network is circumscribed by the interference and loss of path. As we discussed before, algorithms are meant to make decisions about whether or not to choose the optimal BS instead of a nearest one. A nearer BS may be teemed with requests and can be quite disturbing with interference of our best choice BS. Poorer QoE of our users forces us to take that into account [26]. Literature [144] shared the same concern especially when the requested services are executed entirely on the edge servers. How to get better tradeoff results when choosing the right spot remains a difficult issue. In [8], the researchers brought up a new approach to address the inter-interference through coordination.

Their approach is quite appealing but also requires more work to resolve upper layer interference.

Caching process has four phases, and each phase may also be interfering with each other's transmissions on the go. Conventional precaching has become obsolete for its inaccuracy of prediction of popular contents, which brings us back to the designing process. When we are able to make an accurate estimation of content popularity, we then may solve all these, and that still requires a huge amount of effort.

7.3. User Pattern Analysis with Prediction. Based on the analysis of user behavior patterns, parts of the highly popular contents have already been cached and are scattered across the user's equipment so the system can take the joy of transmission without losing anything [145]. The preferences for popularity and prediction results are to be adaptive and flexible. And future models with optimization should be built based on it. Social connections with others also need to be addressed when designing novel caching strategies [51]. The characteristics of the mobile users' behaviors in the network should be attached with enough importance and might be the next hot topic in the fields of edge caching.

Researchers in literature [104] underlined the significance of edge caching in improving the efficiency of content distribution. They put emphasis on mobility issues across the network architecture and lately managed to propose a novel vehicle-aided caching scheme to alleviate pressure coming from BSs in 5G networks. They have done an illustrative job and paved the way for other studies in this field.

There have also been some cases where big data analysis keeps making the wrong choices. For instance, old users sometimes get a higher price than those new users regarding

the same purchasable item. How to make our model remain neutral when serving different types of users with different labels on them is still a challenging issue [146].

7.4. Energy Consumption Reduction in Reality. 5G BSs have significantly higher energy consumption with some performance enhancement compared to 4G, let alone a much larger fluctuation of mobile device users with higher demand for network traffic. A novel and collaborative sleep algorithm developed in [147] can bring down the energy consumption during low load periods. Network resources for computing, storage, and communication are finite on the edge nodes and central server as well. How to keep our network architecture making good use of resources efficiently during peak hours requires further studies. Proactive algorithms deployed in edge caching schemes can cut our energy consumption via a much more efficient caching method.

Battery capacity in mobile users' devices gradually increases with a rising anxiety about battery life among people [148]. Data exchanges in a quicker but more costly way. The effectiveness of wireless transmission also needs another advance in bettering QoE among users. Literature [149] proposed an overview of five aspects in developing a sustainable green 5G network. Higher communication efficiency can be achieved through network traffic modeling and resource allocation, whereas content popularity and channel monitoring make our current 5G and beyond network more sustainable.

7.5. Lessons Learned. The integration of various content placement strategies has been proven by numerous works to alleviate the cost of download and backhaul by roughly one-third, compared to these uncooperative ones. Nevertheless, the computation complexity might be terrifying due to the tremendous amount of content and that is causing our network to be poisoned with redundant data, deteriorating its QoE.

Security issues and interference from others are detrimental to our KPIs, while user pattern analysis addresses both QoS and QoE issues. Some fundamental changes need to be made to exploit the utmost potential of our system. And we believe there still remains much work to do in this field.

8. Conclusion

Edge caching is no novel thing, and in general, the cache placement problem has been thoroughly studied by the research community. Many studies have deployed fancy schemes to simulate the whole caching process with enough attention attached to.

And we can see more and more emphasis on QoE, which is our next step forward towards modernization in the network system as people are more demanding and paying more attention to their own perception of provided service than before. QoE itself has been proven beneficial to mobile edge caching by numerous works. What is more, QoE also urges us to take a big step forward in caching techniques. Our new digital toys feeding on the network are draining our available resources, and relevant countermeasures must be upgraded and reinforced.

In our work, we have provided a comprehensive survey on mobile edge caching and QoE issues together. We began with an extensive introduction of network evolution and then the MEC and mobile edge caching. Related works and studies have been summarized, and we reviewed the history and development of caching schemes. QoE and QoS play as stalwarts in our work and were reviewed followed. We also outlined the merits brought by MEC and edge caching, followed with different issues concerning QoE in the process of edge caching. What is more, several novel applications which are aimed at further advancing our QoE have been introduced and elaborated in detail. The challenges of QoE have been outlined, and we close our paper with fields requiring further investigation.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Conceptualization, methodology, and investigation were contributed by Y.Z. and W.Z. Writing—original draft preparation was contributed by Y.Z. Writing—review and editing was contributed by L.Z. and W.C. Supervision was contributed by W.Z. Wei Zhang is the corresponding author.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China under Grant No. 61802233 and the Shandong Provincial Natural Science Foundation of China under Grant No. ZR2019LZH013 and Grant No. ZR2018MF010. Thanks are due to Qilu University of Technology (Shandong Academy of Sciences) Science, Education and Industry Integration Innovation Pilot Project "Supercomputer Internet Key Technology Research and Application Demonstration."

References

[1] P. Mach and Z. Becvar, "Mobile edge computing: a survey on architecture and computation offloading," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.

[2] C. V. N. Index, "Cisco visual networking index: global mobile data traffic forecast update, 2014–2019," *Technical Report*, 2015.

[3] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: a comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[4] M. Satyanarayanan, "Mobile computing: the next decade," in *Proceedings of the 1st ACM workshop on mobile cloud computing & services: social networks and beyond, ACM*, p. 5, San Francisco California, 2010.

[5] E. Zeydan, E. Bastug, M. Bennis et al., "Big data caching for networking: moving from cloud to edge," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 36–42, 2016.

[6] B. Perabathini, E. Baştuğ, M. Kountouris, M. Debbah, and A. Conte, "Caching at the edge: a green perspective for 5G networks," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, pp. 2830–2835, London, UK, 2015.

[7] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—a key technology towards 5G," *ETSI white paper*, vol. 11, no. 11, pp. 1–16, 2015.

[8] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative mobile edge computing in 5G networks: new paradigms, scenarios, and challenges," *IEEE Communications Magazine*, vol. 55, no. 4, pp. 54–61, 2017.

[9] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5G systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 131–139, 2014.

[10] J. G. Andrews, S. Buzzi, W. Choi et al., "What will 5G be?," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[11] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, 2014.

[12] S. Anokye, S. Mohammed, and S. Guolin, "A survey on machine learning based proactive caching," *ZTE Communications*, vol. 17, no. 4, pp. 46–55, 2020.

[13] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A survey on mobile edge networks: convergence of computing, caching and communications," *IEEE Access*, vol. 5, pp. 6757–6779, 2017.

[14] S. Safavat, N. N. Sapavath, and D. B. Rawat, "Recent advances in mobile edge computing and content caching," *Digital Communications and Networks*, vol. 6, no. 2, pp. 189–194, 2020.

[15] H. S. Goian, O. Y. Al-Jarrah, S. Muhaidat, Y. Al-Hammadi, P. Yoo, and M. Dianati, "Popularity-based video caching techniques for cache-enabled networks: a survey," *IEEE Access*, vol. 7, pp. 27699–27719, 2019.

[16] B. Jedari, G. Premsankar, G. Illahi, M. Di Francesco, A. Mehrabi, and A. Yla-Jaaski, "Video caching, analytics, and delivery at the wireless edge: a survey and future directions," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 431–471, 2021.

[17] J. Shuja, K. Bilal, W. Alasmary, H. Sinky, and E. Alanazi, "Applying machine learning techniques for caching in next-generation edge networks: a comprehensive survey," *Journal of Network and Computer Applications*, vol. 181, article 103005, 2021.

[18] H. Jin, D. Xu, C. Zhao, and D. Liang, "Information-centric mobile caching network frameworks and caching optimization: a survey," *EURASIP Journal on Wireless Communications and Networking*, vol. 2017, no. 1, Article ID 33, 2017.

[19] R. A. Dziyauddin, D. Niyato, N. C. Luong, M. A. M. Izhar, M. Hadhari, and S. Daud, "Computation offloading and content caching delivery in vehicular edge computing: a survey," 2019, http://arxiv.org/abs/1912.07803.

[20] Z. Li, X. Zhou, and Y. Qin, "A survey of mobile edge computing in the industrial Internet," in *2019 7th International Conference on Information, Communication and Networks (ICICN)*, pp. 94–98, Macao, China, 2019.

[21] A. Parvez, I. Rahmati, A. I. Guvenc, H. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3098–3130, 2018.

[22] M. Yamamoto, "A survey of caching networks in content oriented networks," *IEICE Transactions on Communications*, vol. E99.B, no. 5, pp. 961–973, 2016.

[23] Q. Duan, "Cloud service performance evaluation: status, challenges, and opportunities - a survey from the system modeling perspective," *Digital Communications and Networks*, vol. 3, no. 2, pp. 101–111, 2017.

[24] K. M. Morgado, S. Huq, S. Mumtaz, and J. Rodriguez, "A survey of 5G technologies: regulatory, standardization and industrial perspectives," *Digital Communications and Networks*, vol. 4, no. 2, pp. 87–97, 2018.

[25] K. Sha, T. A. Yang, W. Wei, and S. Davari, "A survey of edge computing-based designs for IoT security," *Digital 1580 Communications and Networks*, vol. 6, no. 2, pp. 195–202, 2020.

[26] J. Yao, T. Han, and N. Ansari, "On mobile edge caching," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2525–2553, 2019.

[27] F. Qian, K. S. Quah, J. Huang et al., "Web caching on smart phones: ideal vs. reality," in *Proceedings of the 10th international conference on Mobile systems, applications, and services*, pp. 127–140, Low Wood Bay Lake District UK, 2012.

[28] M. Zhang, H. Luo, and H. Zhang, "A survey of caching mechanisms in information-centric networking," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1473–1499, 2015.

[29] G. Zhang, Y. Li, and T. Lin, "Caching in information centric networking: a survey," *Computer Networks*, vol. 57, no. 16, pp. 3128–3141, 2013.

[30] G I P Association, *5G Vision-the 5G Infrastructure Public Private Partnership: The Next Generation of Communication Networks and Services*, White Paper, 2015.

[31] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: technical misconceptions and business barriers," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 16–22, 2016.

[32] X. Cao, F. Wang, J. Xu, R. Zhang, and S. Cui, "Joint computation and communication cooperation for energy efficient mobile edge computing," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4188–4200, 2019.

[33] E. Baştuğ, M. Bennis, E. Zeydan et al., "Big data meets telcos: a proactive caching perspective," *Journal of Communications and Networks*, vol. 17, no. 6, pp. 549–557, 2015.

[34] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: the role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, 2014.

[35] S. Li, J. Xu, M. Van Der Schaar, and W. Li, "Popularity driven content caching," in *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pp. 1–9, San Francisco, CA, USA, 2016.

[36] Q. Tang, R. Xie, T. Huang, and Y. Liu, "Jointly caching and computation resource allocation for mobile edge networks," *IET Networks*, vol. 8, no. 5, pp. 329–338, 2019.

[37] N. Vance, D. Zhang, and W. D. Edgecache, "Edgecache: a game-theoretic edge-based content caching system for crowd video sharing," in *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems*

(HPCC/SmartCity/DSS), pp. 750–757, Zhangjiajie, China, 2019.

[38] Z. Xie and W. Chen, "Storage-efficient edge caching with asynchronous user requests," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 229–241, 2020.

[39] W. Tang, X. Zhao, W. Rafique, L. Qi, W. Dou, and Q. Ni, "An offloading method using decentralized P2P-enabled mobile edge servers in edge computing," *Journal of Systems Architecture*, vol. 94, pp. 1–13, 2019.

[40] J. Tang, H. Tang, X. Zhang et al., "Energy minimization in D2D-assisted cache-enabled Internet of Things: a deep reinforcement learning approach," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5412–5423, 2020.

[41] X. Zhang and J. Wang, "Heterogeneous statistical QoSdriven power allocation for collaborative D2D caching over edge-computing networks," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pp. 944–953, Dallas, TX, USA, 2019.

[42] C. Fan, T. Zhang, Y. Liu, and Z. Zeng, "Backhaul aware analysis of cache-enabled heterogeneous networks," in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications 1650 (PIMRC)*, pp. 1–7, Istanbul, Turkey, 2019.

[43] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: design aspects, challenges, and future directions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 22–28, 2016.

[44] X. Deng, J. Li, E. Liu, and H. Zhang, "Task allocation algorithm and optimization model on edge collaboration," *Journal of Systems Architecture*, vol. 110, article 101778, 2020.

[45] J. Liu, D. Li, and Y. Xu, "Collaborative online edge caching with Bayesian clustering in wireless networks," *IEEE Internet of Things Journal*, vol. 7, no. 2, pp. 1548–1560, 2020.

[46] Y. Zhang, K. Zhang, J. Cao, H. Liu, and S. Maharjan, "Deep reinforcement learning for social-aware edge computing and caching in urban informatics," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5467–5477, 2020.

[47] S. Mehrizi, A. Tsakmalis, S. Chatzinotas, and B. Ottersten, "A feature-based Bayesian method for content popularity prediction in edge-caching networks," 2019, http://arxiv.org/abs/1905.09824.

[48] Z. Chen, Q. He, L. Liu, D. Lan, H.-M. Chung, and Z. Mao, "An artificial intelligence perspective on mobile edge computing," in *2019 IEEE International Conference on Smart Internet of Things (SmartIoT)*, pp. 100–106, Tianjin, China, 2019.

[49] J. Yao and N. Ansari, "Qos-aware rechargeable UAV trajectory optimization for sensing service," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*, pp. 1–6, Shanghai, China, 2019.

[50] N. Gao, X. Xu, Y. Hou, and L. Gao, "A mobility-aware proactive caching strategy in heterogeneous ultradense networks," in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1–7, Istanbul, Turkey, 2019.

[51] L. Li, C. A. Chan, S. Erfani, and C. Leckie, "Adaptive edge caching based on popularity and prediction for mobile networks," in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, Budapest, Hungary, 2019.

[52] L. Ji and S. Guo, "Energy-Efficient cooperative resource allocation in wireless powered mobile edge computing," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4744–4754, 2019.

[53] M. Mehrabi, D. You, V. Latzko, H. Salah, M. Reisslein, and F. H. Fitzek, "Device-enhanced MEC: multi-access edge computing (MEC) aided by end device computation and caching: a survey," *IEEE Access*, vol. 7, pp. 166079–166108, 2019.

[54] C. Fang, C. Liu, S. Xu et al., "An edge cache-based power-efficient content delivery scheme in mobile wireless networks," in *2019 19th International Symposium on Communications and Information Technologies (ISCIT)*, pp. 294–299, Ho Chi Minh City, Vietnam, 2019.

[55] D. Deng and J. Xia, "Cache-enabled cooperative edge networks for intelligent connected vehicles," *IEEE Access*, vol. 7, pp. 166939–166949, 2019.

[56] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-todevice caching networks: basic principles and system performance," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 176–189, 2015.

[57] H. Rifai, S. Mohammed, and A. Mellouk, "A brief synthesis of QoS-QoE methodologies," in *2011 10th International Symposium on Programming and Systems*, pp. 32–38, Algiers, Algeria, 2011.

[58] S. Khorsandroo, R. Md Noor, and S. Khorsandroo, "A generic quantitative relationship to assess interdependency of QoE and QoS," *KSII Transactions on Internet & Information Systems*, vol. 7, no. 2, pp. 327–346, 2013.

[59] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: a tutorial on video quality assessment," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 1126–1165, 2014.

[60] W.-J. Kim, K.-N. Joo, and C.-H. Youn, "Short-term timevarying request model based chunk caching scheme for live streaming in mobile edge-cloud environment," *IEEE Access*, vol. 7, pp. 177148–177163, 2019.

[61] B. Guo, X. Zhang, Y. Wang, and H. Yang, "Deep-Q network-based multimedia multi-service QoS optimization for mobile edge computing systems," *IEEE Access*, vol. 7, pp. 160961–160972, 2019.

[62] S. Math, L. Zhang, S. Kim, and I. Ryoo, "An intelligent real-time traffic control based on mobile edge computing for individual private environment," *Security and Communication Networks*, vol. 2020, 11 pages, 2020.

[63] R. Wang, M. Li, L. Peng, Y. Hu, M. M. Hassan, and A. Alelaiwi, "Cognitive multi-agent empowering mobile edge computing for resource caching and collaboration," *Future Generation Computer Systems*, vol. 102, pp. 66–74, 2020.

[64] L. E. Chatzieleftheriou, M. Karaliopoulos, and I. Koutsopoulos, "Caching-aware recommendations: nudging user preferences towards better caching performance," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pp. 1–9, Atlanta, GA, USA, 2017.

[65] F. Zeng, Y. Chen, L. Yao, and J. Wu, "A novel reputation incentive mechanism and game theory analysis for service caching in software-defined vehicle edge computing," *Peer-to-Peer Networking and Applications*, vol. 14, no. 2, pp. 467–481, 2021.

[66] L. Li and H. Zhang, "Delay optimization strategy for service cache and task offloading in three-tier architecture mobile edge computing system," *IEEE Access*, vol. 8, pp. 170211–170224, 2020.

[67] X. Zhang and Q. Zhu, "Hierarchical caching for statistical QoS guaranteed multimedia transmissions over 5G edge computing mobile wireless networks," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 12–20, 2018.

[68] Z. Xu, L. Zhou, S. C.-K. Chau, W. Liang, Q. Xia, and P. Zhou, "Collaborate or separate? Distributed service caching in mobile edge clouds," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pp. 2066–2075, Toronto, ON, Canada, 2020.

[69] L. Lei, X. Xiong, L. Hou, and K. Zheng, "Collaborative edge caching through service function chaining: architecture and challenges," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 94–102, 2018.

[70] C. Yousefpour, T. Fung, K. Nguyen et al., "All one needs to know about fog computing and related edge computing paradigms: a complete survey," vol. 98, pp. 289–330, 2019.

[71] Z. Hu, Z. Zheng, T. Wang, L. Song, and X. Li, "Caching as a service: small-cell caching mechanism design for service providers," *IEEE Transactions on Wireless Communications*, vol. 15, no. 10, pp. 6992–7004, 2016.

[72] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," *Wireless Communications and Mobile Computing*, vol. 13, no. 18, 2013.

[73] N. Ansari and X. Sun, "Mobile edge computing empowers Internet of Things," *IEICE Transactions on Communications*, vol. E101.B, no. 3, pp. 604–619, 2018.

[74] N. A. Kiani and N. Ansari, "Toward hierarchical mobile edge computing: an auction-based profit maximization approach," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 2082–2091, 2017.

[75] L. Li, G. Zhao, and R. S. Blum, "A survey of caching techniques in cellular networks: research issues and challenges in content placement and delivery strategies," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 1710–1732, 2018.

[76] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: system architectures, key techniques, and open issues," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 2282–2308, 2016.

[77] A. Argyriou and M. Erol-Kantarci, "A heuristic approach for overlay content-caching network design in 5G wireless networks," in *2016 IEEE Symposium on Computers and Communication (ISCC)*, pp. 621–626, Messina, Italy, 2016.

[78] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "In-network caching and content placement in cooperative small cell networks," in *Proceedings of the 1st International Conference on 5G for Ubiquitous Connectivity*, pp. 128–133, Akaslompolo, Finland, 2014.

[79] X. Sun and N. Ansari, "Latency aware workload offloading in the cloudlet network," *IEEE Communications Letters*, vol. 21, no. 7, pp. 1481–1484, 2017.

[80] S. Zhang, P. He, K. Suto, P. Yang, L. Zhao, and X. Shen, "Cooperative edge caching in user-centric clustered mobile networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 8, pp. 1791–1805, 2018.

[81] W. Jiang, G. Feng, S. Qin, and T. Skeie, "Mobile edge computing: a survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450–465, 2017.

[82] J. Yao and N. Ansari, "Reliability-aware fog resource provisioning for deadline-driven IoT services," in *2018 IEEE global communications conference (GLOBECOM)*, pp. 1–6, Abu Dhabi, United Arab Emirates, 2018.

[83] M. D. D. A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis, "A survey on predicting the popularity of web content," *Journal of Internet Services and Applications*, vol. 5, no. 1, p. 8, 2014.

[84] D. Niyato, D. I. Kim, P. Wang, and L. Song, "A novel caching mechanism for Internet of Things (IoT) sensing service with energy harvesting," in *2016 IEEE International Conference on Communications (ICC)*, pp. 1–6, Kuala Lumpur, Malaysia, 2016.

[85] S. Vural, P. Navaratnam, N. Wang, C. Wang, L. Dong, and R. Tafazolli, "In-network caching of Internet-of-Things data," in *2014 IEEE International Conference on Communications (ICC)*, pp. 3185–3190, Sydney, NSW, Australia, 2014.

[86] R. Urgaonkar, S. Wang, T. He, M. Zafer, K. Chan, and K. K. Leung, "Dynamic service migration and workload scheduling in edge-clouds," *Performance Evaluation*, vol. 91, pp. 205–228, 2015.

[87] M. Musolesi, S. Hailes, and C. Mascolo, "An ad hoc mobility model founded on social network theory," in *Proceedings of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*, pp. 20–24, Venice Italy, 2004.

[88] Q. Lv, Y. Qiao, N. Ansari, J. Liu, and J. Yang, "Big data driven hidden Markov model based individual mobility prediction at points of interest," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5204–5216, 2017.

[89] T. Han and N. Ansari, "Network utility aware traffic load balancing in backhaul-constrained cache-enabled small cell networks with hybrid power supplies," *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2819–2832, 2017.

[90] R. Yu, S. Qin, M. Bennis et al., "Enhancing software-defined ran with collaborative caching and scalable video coding," in *2016 IEEE International Conference on Communications 1860 (ICC)*, pp. 1–6, Kuala Lumpur, Malaysia, 2016.

[91] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility aware caching for content-centric wireless networks: modeling and methodology," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 77–83, 2016.

[92] B. Hu, Y. Chen, Z. Huang, N. A. Mehta, and J. Pan, "Intelligent caching algorithms in heterogeneous wireless networks with uncertainty," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS*, pp. 1549–1558, Dallas, TX, USA, 2019.

[93] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, "Backhaul-aware caching placement for wireless networks," in *2015 IEEE global communications conference (GLOBECOM)*, pp. 1–6, San Diego, CA, USA, 2015.

[94] B. Zhou, Y. Cui, and M. Tao, "Stochastic content-centric multicast scheduling for cache-enabled heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6284–6297, 2016.

[95] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.

[96] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Multicast-aware caching for small cell networks," in *2014 IEEE wireless communications and networking conference (WCNC)*, pp. 2300–2305, Istanbul, Turkey, 2014.

[97] B. Zhou, Y. Cui, and M. Tao, "Optimal dynamic multicast scheduling for cache-enabled content-centric wireless networks," *IEEE Transactions on Communications*, vol. 65, no. 7, pp. 2956–2970, 2017.

[98] M. Alreshoodi and J. Woods, "Survey on QoE,QoS correlation models for multimedia services," 2013, http://arxiv.org/abs/1306.0221.

[99] D. Takahashi, V. Hands, and V. Barriac, "Standardization activities in the ITU for a QoE assessment of IPTV," *IEEE Communications Society magazine*, vol. 46, no. 2, pp. 78–84, 2008.

[100] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Transactions on Networking*, vol. 22, no. 5, pp. 1444–1462, 2014.

[101] S. Ioannou and S. Weber, "A survey of caching policies and forwarding mechanisms in information-centric networking," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2847–2886, 2016.

[102] N. Laoutaris, "A closed-form method for LRU replacement under generalized power-law demand," 2007, http://arxiv.org/abs/0705.1970.

[103] J. Gu, W. Wang, A. Huang, H. Shan, and Z. Zhang, "Distributed cache replacement for caching-enable base stations in cellular networks," in *2014 IEEE International Conference on Communications (ICC)*, pp. 2648–2653, Sydney, NSW, Australia, 2014.

[104] K. Zhang, S. Leng, Y. He, S. Maharjan, and Y. Zhang, "Cooperative content caching in 5G networks with mobile edge computing," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 80–87, 2018.

[105] S. Sengupta, R. Amuru, R. M. Tandon, and T. C. Buehrer, "Learning distributed caching strategies in small cell networks," in *2014 11th International Symposium on Wireless Communications Systems*, pp. 917–921, Barcelona, Spain, 2014.

[106] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *2010 Proceedings IEEE INFOCOM, Citeseer*, pp. 1–9, San Diego, CA, USA, 2010.

[107] W. Jiang, G. Feng, and S. Qin, "Optimal cooperative content caching and delivery policy for heterogeneous cellular networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 5, pp. 1382–1393, 2017.

[108] S. Wang, X. Zhang, K. Yang, L. Wang, and W. Wang, "Distributed edge caching scheme considering the tradeoff between the diversity and redundancy of cached content," in *2015 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 1–5, Shenzhen, China, 2015.

[109] B. Chen, C. Yang, and G. Wang, "Cooperative device-to device communications with caching," in *2016 IEEE 83rd vehicular technology conference (VTC spring)*, pp. 1–5, Nanjing, China, 2016.

[110] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation caching and routing algorithms for massive mobile data delivery," in *2013 IEEE global communications conference (GLOBECOM)*, pp. 3534–3539, Atlanta, GA, USA, 2013.

[111] P. Ostovari, A. Khreishah, and J. Wu, "Cache content placement using triangular network coding," in *2013 IEEE wireless communications and networking conference 1855 (WCNC)*, pp. 1375–1380, Shanghai, China, 2013.

[112] Z. Chen and M. Kountouris, "D2D caching vs. small cell caching: where to cache content in a wireless network?," in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–6, Edinburgh, UK, 2016.

[113] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *2015 IEEE international conference on communications (ICC)*, pp. 3358–3363, London, UK, 2015.

[114] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3665–3676, 2014.

[115] H. Wei, H. Luo, Y. Sun, and M. S. Obaidat, "Cache-aware computation offloading in IoT systems," *IEEE Systems Journal*, vol. 14, no. 1, pp. 61–72, 2020.

[116] Y. Lan, X. Wang, D. Wang, Z. Liu, and Y. Zhang, "Task caching, offloading, and resource allocation in D2D-aided fog computing networks," *IEEE Access*, vol. 7, pp. 104876–104891, 2019.

[117] Y. Wu, S. Yao, Y. Yang et al., "Challenges of mobile social device caching," *IEEE Access*, vol. 4, pp. 8938–8947, 2016.

[118] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 907–922, 2016.

[119] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 849–869, 2015.

[120] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4286–4298, 2014.

[121] M. Sheng, W. Han, C. Huang, J. Li, and S. Cui, "Video delivery in heterogenous crans: architectures and strategies," *IEEEWireless Communications*, vol. 22, no. 3, pp. 14–21, 2015.

[122] M. Vaser and S. Forconi, "QoS KPI and QoE KQI relationship for LTE video streaming and VoLTE services," in *2015 9th International Conference on Next Generation Mobile Applications, Services and Technologies*, pp. 318–323, Cambridge, UK, 2015.

[123] S. Aroussi, T. Bouabana-Tebibel, and A. Mellouk, "Empirical QoE/QoS correlation model based on multiple parameters for VoD flows," in *2012 IEEE global communications conference (GLOBECOM)*, pp. 1963–1968, Anaheim, CA, USA, 2012.

[124] G. Li, M. Wang, J. Feng et al., "Understanding user generated content characteristics: a hot-event perspective," in *2011 IEEE International Conference on Communications (ICC)*, pp. 1–5, Kyoto, Japan, 2011.

[125] Y. Zeng, Y. Huang, Z. Liu, and Y. Yang, "Joint online edge caching and load balancing for mobile data offloading in 5G networks," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pp. 923–933, Dallas, TX, USA, 2019.

[126] H. Wang, Y. Wang, R. Sun, S. Guo, and H. Li, "Joint video caching and user association with mobile edge computing," in *2019 IEEE wireless communications and networking conference workshop (WCNCW)*, pp. 1–6, Marrakech, Morocco, 2019.

[127] S. S. Kafiloğlu, G. Gür, and F. Alagöz, "Multidimensional content modeling and caching in D2D edge networks," in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1–6, Istanbul, Turkey, 2019.

[128] M. Sapienza, E. Guardo, M. Cavallo, G. La Torre, G. Leombruno, and O. Tomarchio, "Solving critical events through mobile edge computing: an approach for smart cities," in *2016 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 1–5, St. Louis, MO, USA, 2016.

[129] D. Liu and C. Yang, "Cache-enabled heterogeneous cellular networks: comparison and tradeoffs," in *2016 IEEE International Conference on Communications (ICC)*, pp. 1–6, Kuala Lumpur, Malaysia, 2016.

[130] J. Rao, H. Feng, and Z. Chen, "Exploiting user mobility for D2D assisted wireless caching networks," in *2016 8th International Conference onWireless Communications & Signal Processing (WCSP)*, pp. 1–5, Yangzhou, China, 2016.

[131] J. Yao and N. Ansari, "Energy-aware task allocation for mobile IoT by online reinforcement learning," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*, pp. 1–6, Shanghai, China, 2019.

[132] J. Quevedo, D. Corujo, and R. Aguiar, "A case for ICN usage in IoT environments," in *2014 IEEE global communications conference*, pp. 2770–2775, Austin, TX, USA, 2014.

[133] X. Sun and N. Ansari, "Dynamic resource caching in the IoT application layer for smart cities," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 606–613, 2018.

[134] S. A. R. Zaidi, M. Ghogho, and D. C. McLernon, "Information centric modeling for two-tier cache enabled cellular networks," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, pp. 80–86, London, UK, 2015.

[135] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Content caching at the wireless network edge: a distributed algorithm via belief propagation," in *2016 IEEE International Conference on communications (ICC)*, pp. 1–6, Kuala Lumpur, Malaysia, 2016.

[136] M. Dehghan, A. Seetharam, B. Jiang et al., "On the complexity of optimal routing and content caching in heterogeneous networks," in *2015 IEEE conference on computer communications (INFOCOM)*, pp. 936–944, Hong Kong, China, 2015.

[137] J. Leguay, G. S. Paschos, E. A. Quaglia, and B. Smyth, "CryptoCache: network caching with confidentiality," in *2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, Paris, France, 2017.

[138] C. Kolias, G. Kambourakis, and S. Gritzalis, "Attacks and countermeasures on 802.16: analysis and assessment," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 487–514, 2012.

[139] M. Mukherjee, R. Matam, L. Shu et al., "Security and privacy in fog computing: challenges," *IEEE Access*, vol. 5, pp. 19293–19304, 2017.

[140] D. Kim, J. Bi, A. V. Vasilakos, and I. Yeom, "Security of cached content in NDN," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 2933–2944, 2017.

[141] T. Wang, Z. Zheng, M. H. Rehmani, S. Yao, and Z. Huo, "Privacy preservation in big data from the communication perspective|a survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 753–778, 2018.

[142] J. H. Abawajy, M. I. Ninggal, and T. Herawan, "Privacy preserving social network data publication," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1974–1997, 2016.

[143] S. Zhang, W. Sun, and J. Liu, "An optimized spatially cooperative caching strategy for heterogeneous caching network," in *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pp. 1685–1689, Tangier, Morocco, 2019.

[144] S. Shekhar, A. Chhokra, H. Sun et al., "Urmila: dynamically trading-off fog and edge resources for performance and mobility-aware IoT services," *Journal of Systems Architecture*, vol. 107, article 101710, 2020.

[145] N. Garg, M. Sellathurai, V. Bhatia, B. Bharath, and T. Ratnarajah, "Online content popularity prediction and learning in wireless edge caching," *IEEE Transactions on Communications*, vol. 68, no. 2, pp. 1087–1100, 2020.

[146] T. Haque and N. Abu-Ghazaleh, "Wireless software defined networking: a survey and taxonomy," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2713–2737, 2016.

[147] K.-C. Chang, K.-C. Chu, H.-C. Wang, Y.-C. Lin, and J.-S. Pan, "Energy saving technology of 5G base station based on internet of things collaborative control," *IEEE Access*, vol. 8, pp. 32935–32946, 2020.

[148] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, "An overview of sustainable green 5G networks," *IEEE Wireless Communications*, vol. 24, no. 4, pp. 72–80, 2017.

[149] M. Lauridsen, *Studies on Mobile Terminal Energy Consumption for LTE and Future 5G, Department of Electronic Systems*, Aalborg University, 2015.