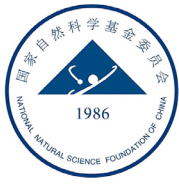


Cross-modal communication technology: A survey

Authors	Wei, Xin;Wu, Dan;Zhou, Liang;Guizani, Mohsen
Citation	X. Wei, D. Wu, L. Zhou, and M. Guizani, "Cross-modal communication technology: A survey," Fundamental Research, vol. 5, no. 5, pp. 2256–2267, Sep. 2025, doi: 10.1016/J.FMRE.2023.08.002
DOI	10.1016/j.fmre.2023.08.002
Publisher	Elsevier
Download date	2026-06-15 04:53:59
Link to Item	https://hdl.handle.net/20.500.14634/1798



Contents lists available at ScienceDirect

Fundamental Research

journal homepage: <http://www.keaipublishing.com/en/journals/fundamental-research/>

Review

Cross-modal communication technology: A survey

Xin Wei^{a,b}, Dan Wu^c, Liang Zhou^{a,b,*}, Mohsen Guizani^d

^a School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

^b Key Lab of Broadband Wireless Communication and Sensor Network Technology, Ministry of Education, Nanjing 210003, China

^c College of Communications Engineering, Army Engineering University of PLA, Nanjing 210007, China

^d Machine Learning Department, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi 999041, UAE

ARTICLE INFO

Article history:

Received 1 March 2023

Received in revised form 2 August 2023

Accepted 16 August 2023

Available online 9 September 2023

Keywords:

Cross-modal communications

Multi-modal service

Semantics

Audio-visual

Haptics

ABSTRACT

In the 5G era and beyond, multi-modal services that integrate audio, visual, and haptic signals are expected to become dominant applications. To support multi-modal services, the concept of cross-modal communications, which involves collaborative audio-visual and haptic interactions, has emerged. Despite significant research about cross-modal communication technology being conducted, a comprehensive literature review on this topic is lacking. To fill this gap, this paper presents a detailed survey on cross-modal communication technology. First, it provides a highly summarized description of representative research attempts in audio-visual and haptic communications, which serve as the foundation for cross-modal communications. Then, it delves into various aspects of cross-modal communications, including architectural, cross-modal coding, cross-modal transmission, cross-modal signal reconstruction, the essence of semantics, and prototype systems. Finally, it discusses conclusions and future research directions. This paper is expected to promote the theoretical research and practical applications of cross-modal communications.

1. Introduction

The convergence of wireless communications and multimedia technologies has led to the emergence of multi-modal services, which are expected to become leading multimedia applications in the 5G era and beyond [1,2]. Examples of these multi-modal services include industrial manipulation, e-health care, remote education, etc. In industrial manipulation, audio-visual and haptic signal feedback can help humans control the teleoperator to fulfil various complicated tasks more accurately. In e-health care, doctors can see and touch a patient's targeted bones or organs to conveniently implement tele-diagnosis or tele-surgery. In remote education, students can be provided with virtual interactive learning resources to improve their learning outcomes and immersive experience during the COVID-19 pandemic. Compared with traditional multimedia services, multi-modal services integrate both audio-visual and haptic modalities, which also have temporal, spatial, and semantic relevance (e.g., time of occurrence, spatial position, and represented content) and can bring human interactive and immersive experience. Therefore, constructing new multimedia communication paradigm and technology to support multi-modal services is necessary.

Historically, research in communications technology has prioritized audio-visual communication, as it has been the dominant mode of communication for existing multimedia services. The acquisition, transmis-

sion, reception, and rendering of audio and visual information have become well-established and mature technologies, with high levels of quality such as high-definition and beyond [3]. In previous audio-visual communications, there are audio/visual coding techniques with high compression ratios, transmission strategies with high efficiency, and receiving and rendering schemes with high fidelity and user experience, some of which have become international standards. Although audio-visual communication can create a convincing sense of presence in a remote environment, it is insufficient for effectively supporting physical interactions and manipulations, which are crucial components of multi-modal services.

To enable physical human-machine and human-human interactions, especially touch perception and feedback in a distant environment, haptic communications have received much attention in recent years, belonging to a representative ultra-reliable and low-latency communications. Unlike audio-visual communications, haptic-enabled teleoperation systems involve bidirectional exchange of haptic signals over a wireless communication network. It also involves humans and closes a global control loop between the humans and the actuators/teleoperators. Therefore, when regarding haptic communications, teleoperation quality and system stability are very sensitive to communication latency [4]. To meet the requirements of ultra-low latency and ultra-high reliability in haptic communications, new methodologies and technolo-

* Corresponding author.

E-mail address: liang.zhou@njupt.edu.cn (L. Zhou).

<https://doi.org/10.1016/j.fmre.2023.08.002>

2667-3258/© 2023 The Authors. Publishing Services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

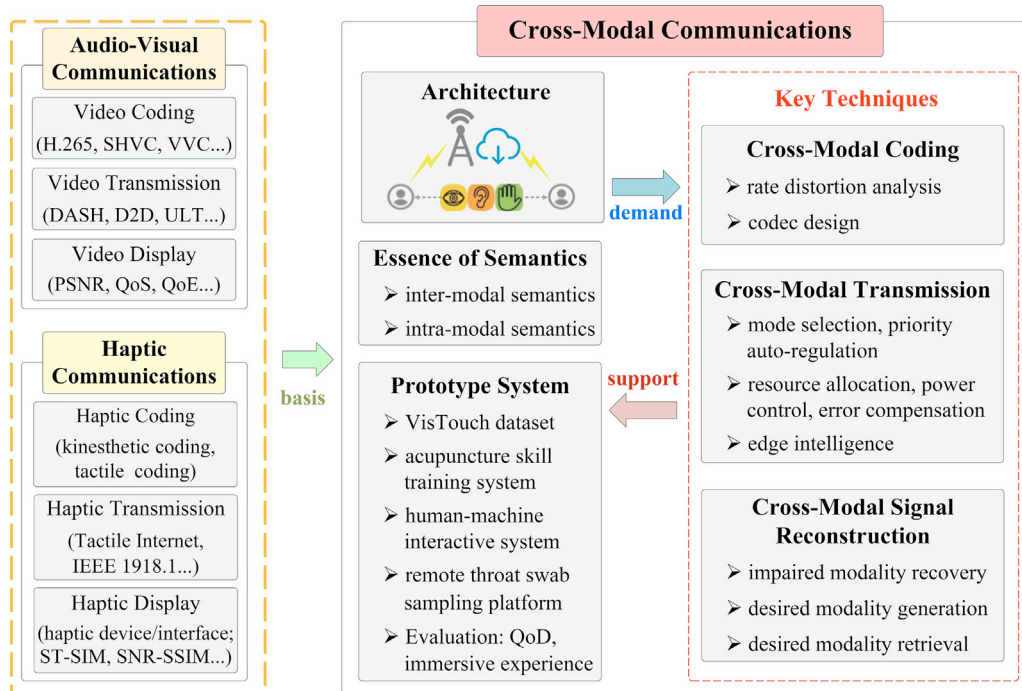


Fig. 1. Organization of this paper.

gies have been developed for the necessary infrastructure, such as the Tactile Internet [5]. The Tactile Internet represents a paradigm shift from content-delivery to skill-set delivery networks, and is capable of adding interactive capabilities to the existing multimedia services.

Audio-visual communications and haptic communications are well suited to audio-visual and haptic signal transmission and processing, respectively. However, they do not affect multi-modal services where audio-visual and haptic signals coexist because there are substantial differences among audio-visual and haptic modalities. Typically, haptic streaming supports an ultra-reliable low latency communications scenario, while audio-visual streaming with a large volume needs adequate bandwidth and belongs to the enhanced mobile broadband communications scenario. To effectively handle the above issues and support multi-modal services, collaborative audio-visual and haptic communications, named cross-modal communications, have been considered as an effective and promising paradigm [6,7]. Essentially, the core idea of cross-modal communications is to thoroughly break barriers and fully leverage potential correlations (or “semantics”) among different modality signals. Taking the material surface detection in industrial manipulation as an example, similar materials in the visual modality have the indistinguishable touch sensories for humans. Motivated by this, redundancy elimination of the haptic modality before transmission can be realized with the aid of visual modality. Moreover, after transmission, the quality of visual signals is prone to unexpected packet loss and delayed arrival of visual streams. Due to the potential correlations among visual and haptic modalities for the same material surface, haptic-aided visual recovery can also be performed. Overall, the ultimate goal of cross-modal communications is to realize ultra-reliable and low-latency communications for haptic streaming as well as high throughput for audio-visual streaming, substantially and considerably improving humans’ immersive experience.

Despite significant progress in audio-visual and haptic communications, a comprehensive literature review of cross-modal communication technology and its relation to existing multimedia communications has yet to be presented. This research gap has

prompted the current paper to provide a detailed review of this topic in response to the rapid development and pressing demands of multi-modal services. This paper makes the following contributions:

We present a summary of existing audio-visual communication and haptic communication research works, which are the origin and basis of cross-modal communications.

We provide an in-depth description of works on cross-modal communications, including architecture, cross-modal coding, cross-modal transmission, and cross-modal signal reconstruction. The description indicates the essence of semantics in cross-modal communications. Moreover, it also provides several developed cross-modal communication prototype systems.

Finally, we give the conclusion and several promising future research directions about this topic.

The organization of this paper is shown in Fig. 1. First, audio-visual communications and haptic communications are introduced in Section 2 and Section 3, respectively. Then, the architecture, key techniques, and prototype systems of cross-modal communications are introduced in Section 4. Finally, the conclusion and future research directions are given in Section 5.

2. Audio-visual communications

During audio-visual communications, video cameras and microphones at source terminals capture ultra-high-definition content and perform efficient audio and visual coding, achieving remarkable compression. Then, the audio-visual streams are allocated appropriate resources and delivered according to the transmission conditions. Finally, high-resolution monitors or virtual reality head-mounted displays at destination terminals provide a high-end auditory-visual experience. Therefore, techniques such as audio-visual related coding, transmission, and display are indispensable. As videos dominate in audio-visual communications, we primarily focus on video communication techniques in the following.

2.1. Video coding

Over the past few decades, researchers have dedicated significant efforts to developing efficient tools for video coding. In 2013, the ITU-T VCEG and ISO/IEC MPEG's Joint Video Team released the H.265/HEVC standard [8]. Compared with its previous generation standard, H.265 particularly concerns issues about the increased video resolution and the increased use of parallel processing architectures, essentially addressing all existing applications of the H.264/AVC standard. However, the original H.265 does not fully consider the adaptation for dynamic environments and flexibility demands for modern video transmission systems. In response to this need, a scalable high-efficiency video coding (SHVC) standard for H.265 was proposed [9]. The term “scalable” refers to the ability to remove parts of the video bitstream to adapt it based on the varying needs of end users, as well as different terminal capabilities or network conditions. For low user demands or bad transmission environment, only part of the bitstream with a baseline profile is delivered, while the other part of the bitstream belonging to high and ultra-high profiles is dropped. Otherwise, all of the bitstream will be transmitted. The most recent international standard, finalized in 2020, is the versatile video coding (VVC) standard, which can achieve a significant reduction in the bit rate (approximately 50% compared to H.265) while maintaining equal video quality [10]. By utilizing new features, VVC provides greater versatility for video with resolutions beyond standard and high-definition, high dynamic range and wide colour gamut video, adaptive streaming with resolution changes, computer-generated and screen-captured video, ultra low-latency streaming, 360-degree immersive video, and more.

2.2. Video transmission

Dynamic adaptive streaming over HTTP (DASH), developed by 3GPP and MPEG, has become one of the most widely-used visual streaming standards [11]. By defining the media presentation description, the segment formats and the control algorithm, DASH can change the video resolution according to the terminal's network capacity. Based on DASH, several improvements have been designed. In ref. [12], the delivery of variable bit rate (VBR) video in on-demand streaming scenarios is examined. VBR delivery can effectively capture the fluctuating characteristics of videos and implement an adaptive algorithm by utilizing the instant bitrates of future segments. A framework called D-DASH, which combines deep learning and reinforcement learning techniques to optimize the performance of DASH, is proposed [13]. This framework can effectively handle the complexity and variability of video content and mobile wireless channels.

In addition, the explosive growth of mobile video services has presented a significant challenge to current cellular networks. To address issues such as limited storage capacity, discrepant computational abilities, dynamic communication environments, random network establishment, and diverse services of large-scale video applications, device-to-device (D2D) communication techniques are considered. An energy-efficient video content delivery system via D2D communication is proposed [14]. By exploring the relationship between coding, storage, and transmission, the proposed system enables large-scale content delivery among mobile devices with constrained energy, unpredictable demand, limited storage, random mobility, and opportunistic transmission.

The video transmission schemes mentioned above are prone to suffer from the cliff effect due to dynamic channel variations. To address this problem, an uncoded (or pseudo-analog) linear-transformed transmission (ULT) strategy is born [15]. This strategy uses continuous numbers instead of discrete bits to represent video signals and employs “analogue-like” amplitude modulation to pursue approximate delivery of information variables, rather than exact delivery. By taking this measure, the ULT greatly reduces the transmission errors originating from channel quality variations.

2.3. Video display

Upon receiving video signals, metrics are utilized to evaluate the quality of the displayed video. The two main conventional metrics for video quality representation are signal quality metrics, such as peak-signal-to-noise-ratio (PSNR), and system quality metrics, such as quality of service (QoS). PSNR reflects the relative strength of noise or distortion in video frames, while QoS mainly concerns factors during transmission. However, these conventional metrics do not consider human-related and context-related factors, and are therefore incapable of representing a viewer's true experience. As a result, quality of experience (QoE) has become an important metric. In ref. [3], a review of selected issues pertaining to QoE and its recent applications in video transmission is presented, including QoE modelling with various influencing factors, QoE assessment during video transmission, and QoE management for video transmission optimization over networks. In ref. [16], the key technologies and realizations of multimedia QoE evaluation are thoroughly investigated. The study indicates that QoE can be predicted from both subjective and objective features, and serves as an effective tool for guiding video communications.

3. Haptic communications

Compared with audio-visual communications, there exist significant distinctions about haptic communications. For example, in bilateral teleoperation systems, haptic signals are exchanged bidirectionally over the network, enabling a global control loop between humans and actuators/teleoperators. These systems transport touch and actuation in real time via a medium supported by the Tactile Internet [5]. The IEEE “Tactile Internet” standards working group, designated IEEE 1918.1, has undertaken pioneering work in this area [17]. However, in order to realize haptic communications and the Tactile Internet, several key issues need to be addressed.

First, considering typical operations in bandwidth-limited networks, techniques for haptic data compression by exploiting the limits of human haptic perception are needed. Moreover, whether the layered approaches in video coding (*i.e.*, SHVC) can be introduced to haptic signals is worth consideration. Second, for haptic transmissions, ultra-responsive connectivity, ultra-reliability, and security are key requirements. To guarantee haptic steering and control of real and virtual objects without creating cybersickness, a roundtrip latency of 1 ms is necessary. Additionally, a fixed-line carrier-grade reliability of seven nines (99.99999%) is critical to ensure packet losses are kept to a minimum and the transparency of the system. Security must be integrated into the physical transmission with a low computational overhead. Third, lightweight and portable haptic devices need further development to allow users to touch, feel, and manipulate objects in both real and virtual environments. Furthermore, the quality of the received haptic signals and the user experience, from both subjective and objective aspects, need to be carefully considered.

3.1. Haptic coding

Haptic data come from either kinaesthetic perception or tactile perception. The former involves force, position, and velocity information sensed by muscles, joints, and tendons of the human body. The latter pertains to surface texture, friction, and other sensations detected by different types of mechanoreceptors in the skin. In kinaesthetic coding, the goal is to reduce data without introducing algorithmic delays due to the strict requirements for transmission stability and low latency. The primary challenge for real-time haptic interaction is how to reduce high haptic packet rates, which is fundamentally different from video compression. The work in ref. [18] proposes two main categories of haptic coding algorithms. The first type is perceptual deadband-based, where the perceptual deadband represents an area below the defined perception threshold. Haptic samples falling within the deadband can be elimi-

nated, as the associated signal change is too small to be perceptible. The second type is predictive coding, where an estimation algorithm is used to predict future haptic sample values from previous data. Only the haptic sample that has more than a barely noticeable difference from the current predicted sample needs to be transmitted. Furthermore, these two kinds of algorithms can be combined to reduce haptic data traffic with little or no influence on the quality of haptic-enabled telepresence interactions [19].

Different from the above works focusing on the compression of kinesthetic information in the context of bilateral teleoperation systems, Hassen et al. [20] makes an effort to perform vibrotactile signal compression. Specifically, it presents PVC-SLP, a touch information coding approach that employs an acceleration sensitivity function. In this scheme, a linear prediction with sparsity constraints is introduced on the residual and the predictor coefficients. Experimental results show that the PVC-SLP perceptually outperforms the competing schemes. It has been selected as the haptic codec in the IEEE 1918.1 standard. Steinbach et al. [21] proposes two coding schemes for tactile signals. One is the waveform-based scheme that transforms original tactile signals to frequency domain (for single-point tactile interaction) or spatial domain (for multi-point tactile scenarios) and then borrows ideas about audio codecs [22]. The other is the parametric representation-based scheme in which the transmitter extracts and sends tactile features (e.g. friction, roughness, or hardness of the object surfaces), while the receiver reproduces the tactile impressions according to these features.

3.2. Haptic transmission

The majority of works in haptic communications are concerned with low latency and high reliability haptic transmission. It can be further divided into three main aspects according to the adopted techniques.

(1) *Edge computing/intelligence*: Reducing the geographical distance between user-terminals and the tactile server through the use of mobile edge clouds and cloudlets is an effective solution to address the unprecedented low latency challenge in haptic transmission. Therefore, edge computing and intelligence are promising techniques for haptic stream forwarding and feedback [5]. Moreover, edge computing can not only save energy through caching and computation offloading, but also relieve security issues by processing original data at terminals or edges.

Specifically, Xu et al. [23] tries to reduce latency by improving the efficiency of edge caching. A hybrid edge caching scheme for the Tactile Internet is proposed based on four representative caching methods (local caching, D2D caching, micro-base station caching, and macro-base station caching). Additionally, a replacement policy is proposed to establish rules for cache eviction based on the popularity of cached files. By utilizing this caching scheme, the reuse of data and content is enhanced, while redundant transmissions can be effectively reduced.

Computing offloading is another way to alleviate congestion and reduce the transmission latency in the Tactile Internet. Xiao and Krunz [24] constructs a fog cooperative computing framework. In this framework, several fog nodes carry on computing and energy costs by cooperating and performing offload forwarding. This scheme can improve the efficiency of power usage and effectively reduce the service response time. In our previous research [25], osmotic computing is selected as a supporting technique to address the issue of computation offloading in the cloud and edge integration. Osmotic computing allows for seamless resource assignment and load balancing, which aligns with the goals of the Tactile Internet.

Efficient routing is also a method to achieve ultra-low latency and ultra-high reliability for establishing haptic communications. Fanibhare et al. [26] proposes a fog-based traffic flow algorithm to effectively manage complex and critical network traffic to reduce extra processing and waiting times at each level of cloud. By using multi-path communications, the traffic flow paths from the master to the slave domains are reduced, leading to the round-trip time falling off.

Moreover, intelligent algorithms can be introduced to form edge intelligence [27]. Mukherjee et al. [28], outline the main challenges to leverage edge intelligence at the master, slave, and network domains of the Tactile Internet. An intelligent communication framework for edge computing is proposed, which consists of two building blocks: prediction and decision-making [29]. This framework maximizes resource utilization efficiency while ensuring latency and reliability constraints. To address the limited resources of edge nodes and end-device privacy concerns, a federated learning algorithm is utilized to guarantee that raw haptic data are not transmitted from their origin. Partial task offloading is considered in the Tactile Internet, where a portion of a task is processed locally at the IoT device and the remaining part is offloaded to the edge server [30]. It is important to note that the repeated interactions are cached at edge nodes closer to end-terminals.

(2) *Resource management*: Transmission resource management has a direct impact on latency, reliability, throughput, and QoS for haptic communication systems. The authors exploit the burstiness of haptic packet arrival to optimize bandwidth reservation in the Tactile Internet [31]. Specifically, the packet arrival process of each user is first classified into high or low traffic states. Then, bandwidth reservation is realized by carefully taking into account the classification errors. Compared with works not aware of burstiness, this scheme can effectively save the bandwidth and guarantee the latency and reliability requirements. Radio resource customization for haptic communications is investigated [32]. Based on radio resource virtualization, a network-wide radio resource slicing strategy is first derived. Through heuristic resource allocation with low complexity, radio resource customization can also be provided.

In addition to the resource management schemes over cellular networks mentioned above, D2D resource management models have also received much attention recently. Tang et al. [33] proposes a “resource-centric” framework that enables joint sharing of communication, computation, and caching resources among mobile edge devices, which generalizes existing D2D resource sharing models and potentially enhances the reliability and intelligence of the Tactile Internet. In ref. [34], the authors propose a cache-enabled D2D-assisted content sharing framework for haptic communications, which can decrease transmission latency by providing timely content delivery. They also model distributed cache and power control as a multiuser content delivery game to maximize individual benefits. This framework can effectively perform resource allocation in D2D-based haptic communications.

(3) *Radio access*: To fulfil the strict requirement of delivering real-time haptic feedback and control within 1 ms latency, radio access is also important and needs careful consideration. A software-defined networking controller can provide the functionality of programming the core and radio access network. However, virtual network functions (VNFs) bring flexibility at the expense of additional packet processing delays. Xiang et al. [35] addresses this issue by developing, implementing, and evaluating a chain-based low-latency VNF implementation, which is a management framework for the distributed service function chains. This proposed scheme executes VNF in a distributed manner, either in the kernel space or user space, based on the processing complexity. It provides an efficient framework for haptic communications [36]. In this framework, users in the network are allowed to share their connectivity, acting as access points for increasing the network capacity. In order to motivate users’ devotion, an incentive and pricing mechanism can be further utilized here.

In addition to meeting the low latency requirement, reliability is also an important concern. High reliability is investigated in a cloud-RAN system through erasure coding and multi-path transmission [37]. The central unit splits the original MAC frame into smaller blocks, encodes them, and transmits them over multiple paths, achieving reliable delivery and a reliability-latency trade-off. Chung et al. [38] proposes a time-division multiplexing passive optical network that provides optical connectivity to end users. This is achieved through the use of multiple remote nodes with simple optical power splitters, supporting

an easy-to-use optical connection anywhere in an optical distributed network.

In addition to the above three aspects, security is another key issue when designing haptic transmissions [39]. To handle this issue, extra control and calculation may be supplied when considering various attacks, affecting the latency performance of the system. Therefore, lightweight encryption algorithms and convenient information exchange schemes should be carefully designed. On the other hand, security for haptic communications may be guaranteed with the help of the other modalities, which is worth exploring in the future.

3.3. Haptic display

Input/output devices to display haptic signals or haptic-related properties can support a human's touch sensation and further interactive experience. Current works aim at developing accurate, real-time, miniaturized haptic display devices or interfaces. For kinaesthetic displays, piezoelectric motors or electromagnetic actuators are usually taken as core components of devices [40]. For tactile displays, the first representative display device was a haptic glove. In ref. [41], a scalable tactile glove was used to identify individual objects, estimate their weights, and explore typical tactile patterns. The glove assembles a sensor array (containing 548 sensors) and can interact with 26 different kinds of objects. To perceptually create relevant tactile cues, Strese et al. [42] designed a tactile computer mouse equipped with a series of actuators. By using this mouse, it can provide the human with major tactile dimensions (hardness, friction, warmth, micro-scopic roughness, macro-scopic roughness) during surface material perception. Yao et al. [43] designs a soft, ultrathin, miniaturized and wireless electro-tactile system that delivers current through the hand to induce tactile sensations as the skin-integrated haptic interface. Sun et al. [44] proposes an augmented tactile-perception and haptic-feedback ring. It contains triboelectric and pyroelectric sensors for tactile and temperature perception, and vibrators and nichrome heaters for vibro- and thermo-haptic feedback. These display devices are promising for a more vivid touch experience in the virtual world and in human-machine interactions.

Moreover, evaluation of the received haptic signal quality is of utmost importance and should be consistent with human perception. In ref. [45], the SSIM index is used to compare the similarity between a reference kinaesthetic signal and a compressed force-feedback signal. Meanwhile, Liu et al. [46] and Hassen and Steinbach [47] focus on evaluating tactile signals. Liu et al. [46] proposes an objective metric that combines the signal-to-noise ratio (SNR) and structural similarity (SSIM) to assess the quality of time-varying vibrotactile data. Hassen and Steinbach [47] introduces the spectral temporal similarity (ST-SIM) that uses a spectral measure and a temporal measure based on the human detection probability function and mean subtracted contrast normalized coefficients, respectively. Experimental results show its effectiveness about haptic display evaluation.

4. Cross-modal communications

4.1. Architecture

Based on existing audio-visual and haptic communication techniques, traditional schemes try to mainly transmit audio-visual-haptic streams through multiplexing. One example of a framework for the adaptive application layer multiplexing is Admux [48]. Admux is designed to adapt to the requirements of multi-modal services and incorporates haptic, visual, and audio data. It provides a synchronization scheme among the different modalities and realizes efficient dynamic bandwidth allocation by using an intelligent multiplexer. Another approach for the application layer multiplexing is proposed, which focuses on teleoperation systems with a multi-modal feedback [49]. In this scheme, haptic signals are given high priority, and a preemptive-resume scheduling strategy is used for audio-visual streaming. The main

aim is to avoid delay-jitter of the received haptic signal and guarantee fluent audio-visual streaming. However, these works treat audio-visual and haptic streams separately, only belonging to multi-modal communications. The availability and efficiency of these communication systems are relatively low due to the significant distinction of characteristics and requirements about audio-visual and haptic modalities. Moreover, in several multi-modal services, audio-visual modality signals dominate and are more important than haptic modality signals. Therefore, always giving haptic modality high priority is not suitable under these circumstances.

Different from previous multi-modal communication schemes, we propose a cross-modal communication paradigm that can better support multi-modal services [6]. It tries to break barriers and fully leverage potential correlations among audio, visual, and haptic modality signals in multi-modal services, enhancing human interactive and immersive experience after satisfying low latency, high reliability, and high throughput requirements. The representative architecture of the cross-modal communications is given in Fig. 2. This architecture mainly consists of a master domain, a slave domain, and a network domain. Specifically, a master domain (receiver) is composed of a human operator and a human system interface. The human operator sends position-orientation commands. The human system interface consists of haptic actuators receiving force-/vibro-/thermo-feedback by haptic devices depicted in Section III-C, as well as audio-visual display devices such as PCs, and mobile terminals. It is noted that several haptic actuators are assembled with operators, such as Geomagic Touch and tactile gloves. A slave domain (transmitter) contains haptic perception sensors installed at teleoperators such as mechanical arms, robots, or virtually synthetic haptic resources. It also contains audio-visual acquisition devices or resources. A network domain serves as the connection between the master and slave domains, providing a wireless communication medium. Devices in the slave domain collect auditory, visual, and touch stimuli from the environment or receive commands from the master domain, which are then compressed through encoding techniques to form multi-modal streams. These streams are transmitted through the network domain via wireless channels to the master domain. Finally, the audio, visual, and haptic signals received at the master domain are decoded, reconstructed, and rendered.

According to diverse proportions and degrees of integration of audio, visual, and haptic signals, multi-modal services can be further divided into three categories. Haptic dominant services include remote industrial manipulation, haptic-enabled telesurgery, etc. Audio-visual dominant services include virtual interactive gaming or movies, holographic museum guides, etc. Equally-important services include online immersive shopping, remote immersive education, etc. Different categories of multi-modal services have distinct demands. For example, in haptic dominant services, the haptic modality has higher priority, while low latency and high reliability requirements should be given more attention. In audio-visual dominant services, high throughput should be considered, while in equally-important services, latency, reliability and throughput factors are all essential. Cross-modal communications attempt to explore and leverage potential correlations among different modality signals to satisfy the requirements of multi-modal services. Therefore, there exist several key techniques when designing cross-modal communication schemes and systems.

First, due to the limited communication resources, the system needs to compress multi-modal signals at the slave domain before delivery. Current multi-modal coding schemes typically employ separate codecs for haptic, visual, and audio streams. For instance, H.265 is used for video coding while [18] is used for haptic coding. However, performance metrics, such as rate distortion of the audio-visual and haptic streams, are greatly affected by this scheme. Therefore, it is important to design a cross-modal coding technique to further compress audio-visual-haptic signals and eliminate redundancy among modalities.

Second, due to the obviously distinct requirements among modalities, the system needs to perform efficient multi-modal streaming.

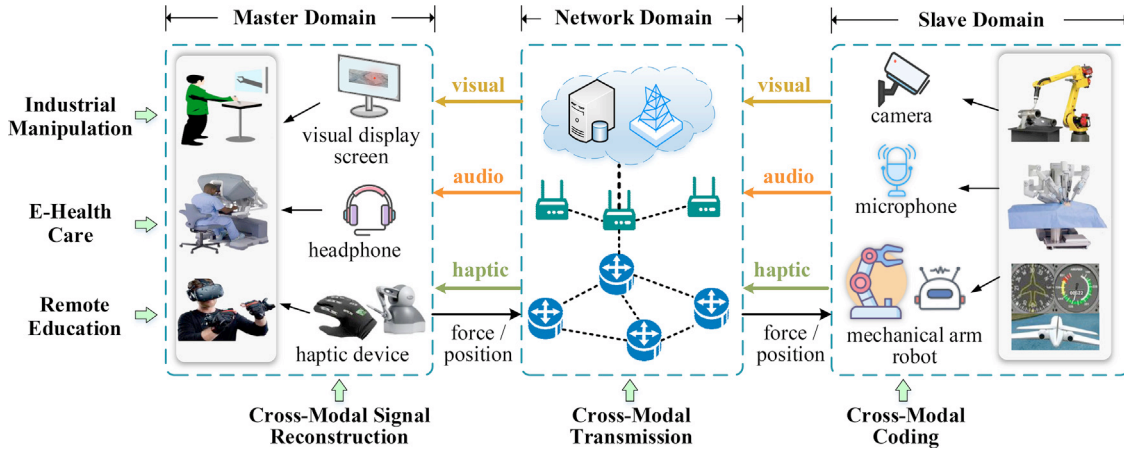


Fig. 2. Architecture of cross-modal communications.

Specifically, haptic streams require low latency and high reliability during transmission, whereas audio-visual streams demand high bandwidth and low jitter. Moreover, for different categories of multi-modal services, the keystone of the corresponding transmission schemes should also be distinguished, especially under dynamic transmission environments. Therefore, constructing a cross-modal transmission framework and scheduling resources are necessary.

Third, due to the transmission impairment and rendering demands, comprehensive signal processing algorithms need to be designed and the human’s immersive experience at the master domain require to be evaluated. In multi-modal services, it can be known that the received audio-visual and haptic signals have potential correlations (*i.e.*, belonging to the same content or having the common property). Therefore, performing a cross-modal signal reconstruction strategy that cleverly extracts these correlations, realizes mutual assistance among modalities, and satisfies immersive experience is worth exploring.

4.2. Cross-modal coding

To compress the amount of multi-modal signals during communications, Zhou and Yuan [50] proposes a cross-modal coding scheme by using the semantic correlation (side information) among different modalities signals. In this scheme, video and haptic signals at the transmitter are separately encoded into video and haptic packets. At the receiver, haptic signals are decoded by the received haptic packets as well as the side information from the received video packets. In other words, side information is explored and utilized to realize cross-modal coding. Moreover, several additional mechanisms such as stability control, hierarchical coding, and cross-modal embedded synchronization can also be applied. During stability control, the haptic arrival latency can be predicted by using channel conditions from video streaming. In hierarchical coding, a scalable coding algorithm is adopted for video signals. In cross-modal embedded synchronization, haptic bits are embedded into the transform results of video frame residuals to guarantee synchronization. Furthermore, Yuan et al. [51] investigates cross-modal coding based on information theory and fully considers side information. It first considers the representation of symbol and semantic relationships among multi-modal sources, defining semantic entropy. Then, the benefit boundary and the minimum number of bits required to compress haptic signals are derived under the rate conditions of video streams. Using this theoretical guidance, a cross-modal codec has been designed that employs AI-enabled cross-modal prediction and channel coding, as shown in Fig. 3. The codec emphasizes the potential correlation between pairs of haptic and video frames. Based on this correlation, the previously decoded video frame is used to estimate the corresponding haptic reference frame during decoding. The experimental results show very

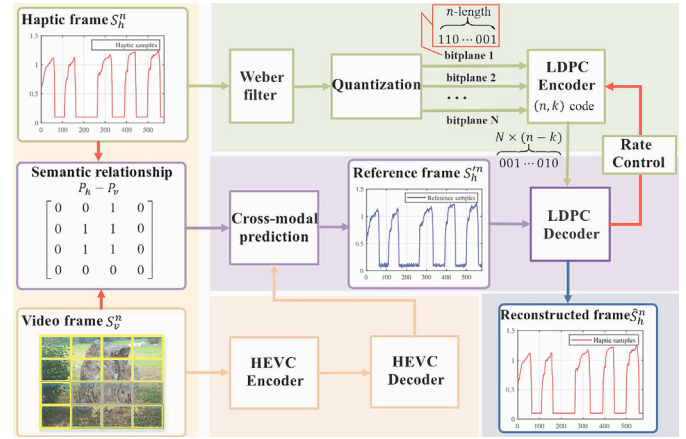


Fig. 3. The framework of a cross-modal codec [51].

impressive coding efficiency and bit error ratio (BER). However, the cross-modal coding technique is still in its infancy. A general codec with compatibility, universality, and low complexity characteristics, close to the theoretical compression limit, must be further investigated.

4.3. Cross-modal transmission

From the above analysis, various service scenarios (*i.e.*, distinguished contents and requests for different categories of multi-modal services), heterogeneous streams (*i.e.*, different transmission requirements for audio-visual and haptic modality streams), and dynamic network environments (*i.e.*, varied channel conditions and limited transmission resources) are three main issues when designing cross-modal transmission strategies.

To address issues regarding distinguished contents and requests in various multi-modal service scenarios, Zhou et al. [52] proposes a cross-modal stream scheduling strategy. First, a hierarchical category framework is constructed that has two motivations. The first is that different applications typically have distinct priorities based on their requests and contents. The other is that consistent with the audio and visual streams, scalable coding can also be deployed for haptic streams. Different from previous schemes with high priority haptic configurations, this framework can flexibly set the priority of audio, visual, and haptic streams. Second, a series of modal-aware scheduling schemes are designed. In terms of haptic-dominant applications, resource multiplexing through spatial diversity and multi-path transmission is deployed to reduce the haptic amount and improve the efficiency of the network resources as much as possible. For audio-visual dominant applications, a network

slicing-based virtual resource is considered, which can be realized by orchestrating, deploying, and arranging the visualized network function. It not only aims to guarantee the communication quality of audio-visual signals with large amount but also tries to reduce the transmission resource occupation of haptic signals on the audio-visual signals. For the equally-important services, mobile edge computing, including unloading decisions and resource allocation, is considered a promising technique. It fully considers the spatio-temporal-semantic relationships among the modal streams and adaptively arranges communication, computing, and caching to achieve the tradeoff among latency, reliability, throughput, and complexity. Third, a scheduling switch strategy for flexible and adaptive scheduling scheme decisions is proposed. It is developed from the semantic smooth perspective for improving the application generality and reducing the performance fluctuation, finally enhancing user experience. Moreover, a dynamic transmission mode selection for multi-modal services is proposed [53]. The combination of three representative transmission modes (cellular, D2D, multiplexing) is taken according to the categories of multi-modal services. Then, the core indicators, such as bandwidth to be predicted for audio-visual streams, latency threshold for haptic streams, are given. Finally, a joint mode selection and resource allocation algorithm is presented, resulting in improvements of throughput, energy utilization, and delay for various multi-modal application scenarios.

To address issues about heterogeneous transmission requirements among modalities, Yang et al. [54] proposes a scheduling scheme for cross-modal transmission. On the one hand, haptic signals are sensitive to transmission latency, which can cause discontinuity in the haptic stream at the receiver. On the other hand, haptic streaming can impact the delivery of audio-visual streams, resulting in reduced quality. As latency and reliability often conflict, a linear model-based prediction mechanism that leverages the correlation within haptic signals is proposed to eliminate haptic stream discontinuity by predicting and sending future haptic signals in advance. Furthermore, a direct D2D link at the receiver's side is established, utilizing the proximity feature of the receivers. This mechanism can compensate for reliability loss resulting from haptic prediction and ensure the transmission quality of the audio-visual stream. The theoretical results are used to develop a minimum power resource consumption search algorithm, obtaining the optimal prediction horizon under the constraints of delay and reliability. By utilizing this cross-modal transmission scheduling framework, wireless resource consumption can be saved without requiring devices to add additional interfaces. An adaptive stream scheduling strategy is proposed, which can not only substantially decrease the stream traffic, but also dramatically reduce the impact of the channel uncertainty and the haptic signals' stochastic arrival on the delay jitter and reliability [55]. Meanwhile, for haptic transmission with higher priority, a joint uplink and downlink resource allocation scheme based on a prediction model is designed. For audio-visual stream transmissions, a power-domain NOMA with LMDC-FEC scheme is presented to fully utilize the remaining limited resources. Different from Yang et al. [54] and Wu and Zhou [55], Wei et al. [56] proposes a cross-modal transmission strategy from the aspects of both reducing packets to be delivered in advance and remedying the lost packets to be received afterwards. The main concept behind this strategy is to create a symbiotic relationship between the visual and haptic modalities. At the transmitter, a scheme is developed to reduce the haptic content by using visual cues. This is accomplished by identifying areas where the haptic and visual contents are similar and only transmitting the necessary haptic frames, thus reducing the burden on the transmission. At the receiver, when the visual packets are not available, degraded or delayed, a fine-grained approach for haptic-to-visual synthesis is used to remedy the situation. It should be noted that this approach does not require the use of paired visual-haptic data for model training, which expands its potential uses. The performance of this strategy is evaluated through experiments on a practical cross-modal communication platform, demonstrating its superiority.

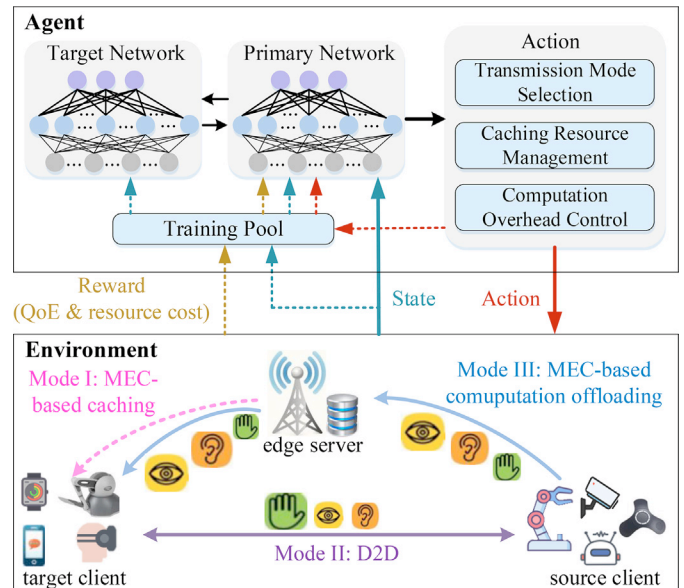


Fig. 4. Reinforcement learning-based cross-modal transmission [7].

To fundamentally address the dynamic network environment problem, artificial intelligence has been introduced in the design of cross-modal transmission strategies. One approach proposed in [7] involves using reinforcement learning to jointly optimize caching, communication, and computation to enable collaborative transmissions of audio, visual, and haptic streams. Specifically, a cloud centre is replaced by edge servers for implementing caching and computing tasks. A hybrid transmission mode, containing mobile edge computing (MEC)-based caching, D2D, and MEC-based computation offloading, is utilized for variable transmission resource conditions. It is noted that for the haptic streams with bursty properties, D2D is recommended with a higher priority due to its characteristics such as low latency, conveniently supporting bidirectional communications. The proposed hybrid transmission modes enable the derivation of caching, communication, and computation costs based on delay, packet loss, jitter, and data rate indicators. To achieve autonomous transmission optimization, deep reinforcement learning or broad reinforcement learning techniques [57] can be applied using the cost functions, as depicted in Fig. 4. An edge intelligence-based cross-modal streaming transmission architecture is proposed for dynamic channel environments [58]. This architecture utilizes the 4C capabilities (caching, computation, communication, control) to the fullest extent. A fast content popularity estimation and clustering algorithm is first provided for content caching. Then, through collaborations by user devices and edge servers, efficient computation offloading can be supported. Subsequently, secure communication techniques such as federated learning and blockchain are considered for protecting personal privacy and data security during transmissions. Finally, an attention-based deep reinforcement learning solution is taken as a control model to formulate the autonomous optimization mechanism for efficient caching, collaborative computing, and secure communication. With this 4C-based autonomous strategy, a desirable cross-modal transmission performance can be achieved.

4.4. Cross-modal signal reconstruction

To guarantee the completeness and quality of the received audio, visual, and haptic signals, cross-modal signal reconstruction techniques are designed by considering potential correlations among modalities. In other words, it fully explores “semantics” within one or two normally-received modality signals and helps reconstruct or generate the remaining but desired modality signals.

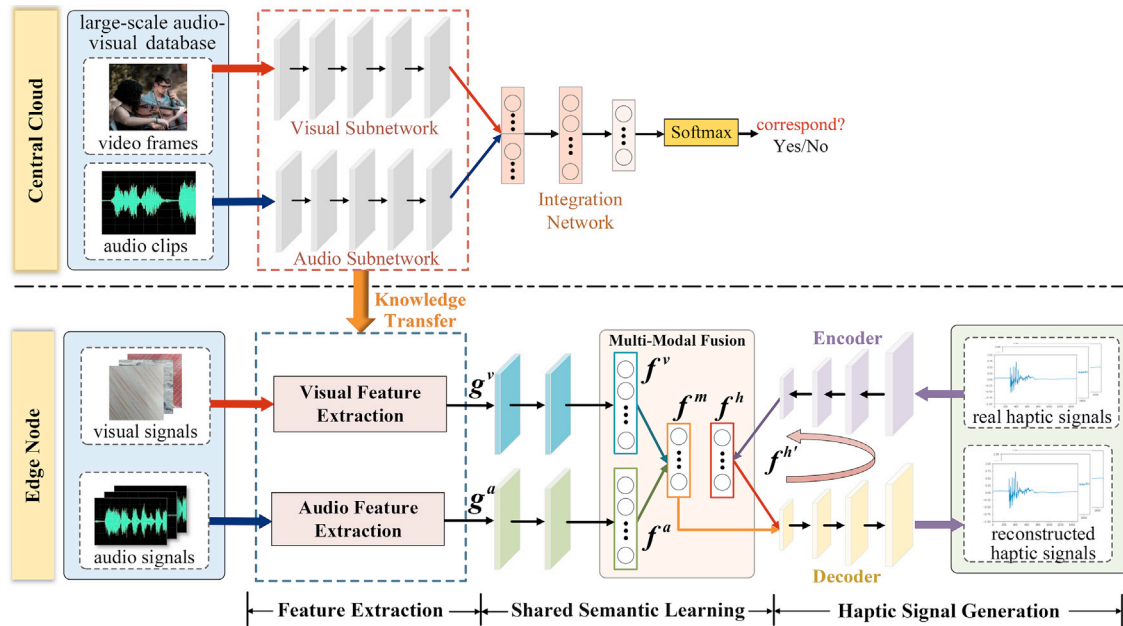


Fig. 5. The proposed audio-visual-aided haptic signal reconstruction approach [59].

Wei et al. [59] analyses two representative scenarios that need cross-modal signal reconstruction techniques. Taking the haptic modality as an example, the appearance of a haptic signal is sudden and unpredictable in several applications, such as industrial control and haptic-enabled telesurgery. It is very sensitive to interference and noise, leading to haptic signal degradation at the receiver. By considering the received audio-visual signals and connecting their potential correlation with the haptic modality, impaired haptic signal reconstruction can be realized. The other scenario is that there are no haptic acquisition devices at the slave domain. Real haptic signals do not exist, but humans at the master domain have touch sensation demands. For example, by introducing haptic information into online immersive shopping or virtual interactive teaching, a multi-dimensional immersive experience can be obtained. Under these circumstances, the desired haptic signals need to be synthesized using the received audio and visual signals, effectively promoting “virtual” touch sensation.

Considering the above demands, Wei and Zhou [7] proposes a transfer learning-based cross-modal comprehensive signal processing paradigm at the receiver. It contains a cross-database knowledge transfer for handling insufficient training samples, cross-modal semantic transfer among modalities for generating a desired modality, and intra-modal characteristic transfer for further recovering or rendering the final signals. Under this paradigm, Wei et al. [59] establishes an audio-visual-aided haptic signal reconstruction (AVHR) approach based on cloud-edge collaborative framework, shown in Fig. 5. Specifically, due to the insufficient received data at the edge node, knowledge can be obtained and transferred from a large-scale audio-visual auxiliary database at the central cloud. The edge node then mines the intrinsic semantic consistency and relevance across modalities, and finally realizes the desired haptic signal reconstruction. It is noted that in this approach, multi-level (signal, feature, and category) constraints are adopted for fine-grained signal processing. An experiment on a constructed practical cross-modal communication platform evaluates the performance of the AVHR.

In ref. [60], an approach for haptic signal reconstruction in cross-modal communications is proposed to address the challenges of complex operations or inefficient feature representations in the literature. The approach involves constructing a force reconstruction network using long short-term memory (LSTM-FRN) to learn mapping correlations from the data distribution, avoiding the need for complex analytical methods. In

addition, a sparse attention module and a metric learning-based constraint are introduced to improve the reconstruction precision and latency of the LSTM-FRN, resulting in a better performance compared to traditional machine learning schemes. Moreover, by leveraging semantics of multi-modal signals, Li et al. [61] presents a universe and robust end-to-end cross-modal signal reconstruction framework. This framework contains a feature extraction module for the source modality signals, a reconstruction module for the desired modality signals, and an evaluation module for guiding further optimization of the other two modules. Regarding reconstruction from video to haptic signals, a 3D convolutional neural network (CNN)-based video extraction subnetwork, a fully convolutional network based generative adversarial network (GAN) generation subnetwork, and a CNN-based GAN discrimination subnetwork are employed for the above three modules. From the experimental results, the framework can achieve high accuracy and reliable reconstruction effects.

Different from existing ideas in cross-modal reconstruction works, Wei et al. [62] proposes a cross-modal signal reconstruction strategy from the perspective of human perceptual facts. The reason is that in the human perception system, inherent interactions among auditory, visual, and haptic sensories also exist, influencing or even constructing a human’s experience. Therefore, the related human perceptual mechanisms can be borrowed during research on cross-modal communications. On the one hand, auditory and haptic sensories, belonging to two representative non-visual perceptions, have many similarities, such as small size, waveforms with little difference in time-frequency domain, and analogous masking effects. This motivates the development of a redundancy elimination mechanism for audio-haptic signals based on time-frequency masking. On the other hand, according to psychological theory, non-visual perceptions can improve visual experiences. Based on this, the visual signals that are impaired or delayed at the receiver can be reconstructed using the unimpaired audio and haptic signals. Therefore, a simple yet effective approach named audio-haptic fused visual signal restoration approach can be designed.

In addition to perform reconstruction from cross-modal signal generation and repairing, retrieval is another method with fast response and high-reliability characteristics. Xu et al. [63] proposes an information recovery technology for cross-modal communications. By retrieving and taking the required data stored at the master domain, signals lost or polluted by noise of a wireless channel during transmission can be recov-

ered. To appropriate the different kinds of multi-modal services and specific scenarios, one-to-one intra-modal retrieval, one-to-one inter-modal retrieval, and one-to-many inter-modal retrieval are designed. Experimental results under various wireless channel conditions show that this scheme has a strong information recovery ability across audio, visual, and haptic signals. Although the premise is that the desired information should be saved or cached at the cloud or edges near the receiving terminals, it can still be taken as a complementary tool for promoting the quality of the received signals.

4.5. The essence of semantics

From the existing works mentioned in above, potential correlations between audio-visual and haptic modalities are vital when designing communication schemes. In other words, these potential correlations can be taken as “semantics” independent of modality, or “inter-modal semantics”. In multi-modal services, semantics can be explored by three main ways. First, meaningful semantic information can be annotated or brought in audio-visual and haptic streams, e.g., category labels of objects in the related video and haptic signals. Under this circumstance, inter-modal semantics can be obtained in a supervised manner. Second, meaningful semantic information cannot be annotated or brought in audio-visual and haptic streams. In other words, inter-semantics should be explored in an unsupervised way. For instance, the delivered haptic packets and audio-visual packets from the same terminal in a network, or the received haptic and audio-visual signals have the same synchronization time stamp, etc. This information can be utilized for exploring semantics. Third, even though the audio-visual and haptic signals at the edges or terminal devices have no semantics to leverage, useful information may exist at other parts in the network, i.e., at the central cloud. Therefore, part of the important inter-semantics can also be obtained and transferred from these nodes by artificial intelligent techniques, such as transfer learning and knowledge distillation.

Owing to the exploration of inter-modal semantics, one modality content can be effectively compressed from the other modalities when designing cross-modal coding mechanisms. Subsequently, one modality stream can be efficiently scheduled from the other modalities when constructing cross-modal transmission strategies. Finally, one modality signal can be realistically generated or repaired from the other modalities when designing cross-modal signal reconstruction algorithms.

On the other hand, with the advancements in 5G, micro-electronics, and AI, semantic communications have been proposed as a more intelligent communication paradigm that focuses on the meaning of transmitted messages rather than just accurate transmission of bit streams [65,66]. Compared with bit-based communications, there are five main characteristics about semantic communications. i) The source is able to extract and encode semantics of a raw message for transmission. The destination should be able to “understand” and infer the message from the received semantics. ii) There are two different types of noises: physical channel noise and semantic noise. It is noted that semantic noise may seriously affect the performance of semantic communications, as ambiguity exists in words, sentences or symbols between extracting and interpreting messages at the source and destination. iii) A semantic communication system is based on knowledge. Background knowledge bases can be established by self-learning at the source and destination. iv) There are diverse metrics to measure the performance of semantic communication systems according to types of original messages or knowledge bases (i.e., text, speech, image, video) [67]. v) Semantic communications can support either full data reconstruction or direct task execution based on the demands of services.

According to the modalities concerned, previous works focus on text, speech, visual-related semantic communications. Specifically, Xie and Qin [68] develops a lightweight distributed semantic communication system for text transmission. Weng and Qin [69] proposes a squeeze-and-excitation network with attention mechanisms for speech transmission. Jankowski et al. [70] proposes an image retrieval over wireless

channels scheme, while [71] presents the first work on joint source-channel coding for video signals over wireless channels. Essentially, “meanings” obtained from respective modality signals, or named “intra-modal semantics”, are utilized for executing semantic communications.

In general, semantic communications have several prominent advantages. It can effectively compress original messages by filtering out the useless, irrelevant, and unessential information and preserving meaningful information [72]. In addition, a semantic communication system is more robust to terrible channel environments such as low SNR region, limited bandwidth, and high BER/sentence error rate (SER). However, in semantic communications, ensuring high reliability remains a significant challenge due to the existence of polysemy and ambiguity issues. After introducing “inter-modal semantics” in cross-modal communications, the hidden but true meanings can be explored in addition to “intra-modal semantics” or explicit semantics. In other words, semantics in cross-modal communications can effectively address the polysemy and ambiguity problems in semantic communications [73].

4.6. Prototype system

Based on theoretical and technical investigations, several cross-modal communication prototype systems have been developed, as shown in Fig. 6 and listed as follows:

Large-scale cross-modal-related dataset [61]: To meet the research needs of cross-modal communications, a platform has been created to collect a large-scale object surface material dataset, called VisTouch (<http://8.133.175.194>). In the master domain, Geomagic Touch is utilized to send control instructions and receive force-feedback haptic signals, while a notebook is used for displaying received videos. In the slave domain, a 4K HDMI camera, a microphone, and a UR3 robotic hand with TeckScan thin-film pressure sensor are employed to touch object surfaces with various materials. The VisTouch dataset contains representative materials in approximately 11 micro-categories (plastic, metal, wood, paper, ceramic, rubber, natural textile, synthetic textile, glass, leather, slate) and 300 micro-categories. It has 800 groups of audio-visual-haptic signals, while the whole amount of data is ten million levels. As audio-visual and haptic signals with potential correlations have common material characteristics, VisTouch is considered to be suitable for designing and evaluating cross-modal communication schemes.

Visual-haptic human-machine interactive system: Based on straight-line servo drives with small profiles and large torques, a glove with haptic perception has been developed for remote human-machine interactions by the authors’ research group. It can follow the trace of the human hand’s movement and receive haptic force feedback from the robotic hand. Specifically, when a human’s hand movement appears, it can trace and record and then transmit the movement data to the Aubo-i3 mechanical arm. According to the movement instructions, the Inspire robotic hand is controlled (i.e., grip the desired object). At the same time, the Inspire robotic hand returns force information to the glove via a wireless channel, enabling a person to perceive real haptic feedback. Moreover, an additional Kinect camera collects the corresponding video signals and compensates for haptic signal impairment due to wireless transmission or mechanical vibration. By using a cross-modal signal reconstruction technique, it can further enhance the reliability of the human-machine interaction. It is noted that this system can play an important role in scenarios about field remote rescue, remote industrial management, etc.

Virtual acupuncture skill training application [60]: Currently, implementing remote practical teaching is still inconvenient, especially for courses needing actual operations or lab experiments. Considering this challenge, we take the acupuncture skill training as an example and construct a virtual interactive application by resorting to cross-modal communication technology. This application involves three components: vivid haptic rendering, augmented reality, and a skill assessment subsystem. Importantly, the vivid haptic rendering includes a cross-modal hap-



Fig. 6. Illustrations of the developed prototype systems: a Large-scale cross-modal communication-related dataset; b Visual-haptic human-machine interactive system; c Virtual interactive acupuncture skill training application [60]; d Remote throat swab sampling platform [64].

tic reconstruction algorithm for synthesizing haptic force feedback and a self-assembled touch sensation perception device (Geomagic Touch connected with an acupuncture needle) for rendering this haptic force feedback. The skill assessment subsystem evaluates students’ manipulation by comparing the needle motion and haptic signals of their training manipulation to the standard manipulation in terms of four aspects: position, velocity, acceleration, and force. After a semester of application, this system can vividly show sensations such as “the qi” and “needling”, correcting and promoting students’ acupuncture skills.

Remote throat swab sampling platform [64]: To facilitate safe and effective COVID-19 diagnosis, a contactless throat swab sampling platform has been developed using cross-modal communication technology. This platform allows for the remote operation of a mechanical arm for precise swab sampling and collection via high-quality visual navigation and haptic feedback. The process involves four stages: cotton swab grasping,

mouth tracking, secretion sampling, and sample collection. To enhance haptic fidelity and visual quality, a semantic-aided cross-modal reconstruction and a user experience-driven stream scheduling strategy have been proposed. The platform achieves an average sampling accuracy of 98% and provides an immersive experience by transmitting and processing video and haptic signals in a cross-modal way. The development of this platform is particularly significant for protecting medical staff and breaking the chains of virus spread.

Finally, evaluation of cross-modal communication systems with the related schemes is also important. Existing works propose evaluation metrics from either human-centric or machine-centric perspectives. For human-centric evaluation, Gao et al. [74] takes the immersive experience (IE) as a metric for the evaluation of multi-modal services equipped with interactive characteristics. It differentiates IE from traditional QoE and concludes key factors influencing IE, which are user-

aware, device-aware, network-aware, context-aware influencing factors. To satisfy timely IE evaluation under large-scale and high-dimensional data, a lightweight approach is also proposed based on multi-view machine learning technique. Furthermore, Gao et al. [75] systematically addresses three fundamental problems about IE from theoretical aspects: which factors influence IE, how to online improve IE, and to what extent of the corresponding IE improvement can be achieved. Therefore, this work can guide efforts for improving IE when designing cross-modal communication schemes. For machine-centric evaluation, quality of decision (QoD), focusing on task accomplishment rather than user satisfaction or data fidelity, was first proposed in [76]. QoD can be calculated by quality estimation from four layers of a specific multi-modal service (environment, sensing, network, and computing layers). In this way, it can effectively evaluate the performance of multi-modal services emphasizing machines instead of humans. Therefore, this metric is able to independently monitor the quality of the acquired, delivered, and processed data from machine decision making aspects. Overall, we think IE and QoD can play complementary roles in comprehensively and precisely evaluating the quality of cross-modal communication systems [77].

5. Conclusion and future research directions

Through this paper, we intend to contribute to the growing area of research in cross-modal communications. Therefore, we first review the related works of audio-visual and haptic communications, which are the origin and basis of cross-modal communications. Then, we provide a comprehensive description of the cross-modal coding, cross-modal transmission, and cross-modal signal reconstruction, which are core components of cross-modal communications. Next, we discuss the essence of semantics that connects cross-modal communications with existing semantic communication schemes and may further enhance communication performance. Finally, we describe several developed cross-modal communication prototype systems. Additionally, inspired by recent works, the present study leaves several open and interesting questions.

First, although the essence of semantics in cross-modal communications has been discussed, further studies need to be conducted. From fundamental theory aspects, future research should quantitatively describe semantics and its influencing effects on polysemy and ambiguity. They should also determine the limits of the multi-modal streaming rate under any semantic distortion as well as the limits of semantic deviation under any multi-modal streaming rate. From the key technology aspect, specific coding schemes combining intra- and inter-modal semantics do not appear in the literature. Under circumstances such as extreme compression and insufficient prior knowledge, how to leverage potential correlations among modalities to reduce polysemy and ambiguity in semantic communications should be examined.

Second, current multi-modal services are mainly concerned with audio, visual, and haptic signals. There are two other important types of sensorial information: olfactory and gustatory. To include olfactory and gustatory information, the existing cross-modal codings, transmission, and signal reconstruction schemes need to be extended or redesigned. Moreover, most previous works explore semantics among modalities by resorting to signal processing and AI techniques. In neuroscience, there exist natural but inherent interactions among a human's auditory, visual, haptic, olfactory and gustatory sensory subsystems that comprehensively determine the experience. Therefore, if these interactions can be fully exploited, they can be transferred and will further improve the theoretical and application research on cross-modal communications.

Third, cross-modal communication technologies may guide the handling of key issues in other fields. For example, the online-offline hybrid instruction paradigm has become mainstream in the area of smart education. In this paradigm, knowledge can be delivered either from a learning platform in online scenarios or from a teacher in an offline

classroom to students, while students can also generate learning feedback (i.e., learning behaviour, testing results). The interactions in the online and offline instruction scenarios can be considered two different modality streamings due to their substantially distinguished characteristics. Previous works perform them separately, and an information exchange barrier exists. Therefore, resorting to the core ideas of cross-modal communications may break this barrier, realize deep integration of online and offline instruction, and ultimately enhance students' learning achievements and experiences.

Declaration of competing interest

The authors declare that they have no conflicts of interest in this work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62231017, 62071254, 62122094, 62277032), the Qinglan Project of Jiangsu Province, the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

- [1] A. Covaci, L. Zou, I. Tal, et al., Is multimedia multisensorial? - A review of multimedia systems, *ACM Comput. Surv.* 51 (5) (2019) 91.
- [2] X. Shen, J. Gao, M. Li, et al., Toward immersive communications in 6G, *Front. Comput. Sci.* 4 (2023) 106848.
- [3] T. Zhao, Q. Liu, C.W. Chen, QoE in video transmission: A user experience-driven strategy, *IEEE Commun. Surv. Tut.* 19 (1) (2017) 285–302.
- [4] K. Antonakoglou, X. Xu, E. Steinbach, et al., Toward haptic communications over the 5G tactile internet, *IEEE Commun. Surv. Tut.* 20 (4) (2018) 3034–3059.
- [5] M. Simsek, A. Aijaz, M. Dohler, et al., 5G-enabled tactile internet, *IEEE J. Sel. Areas Commun.* 34 (3) (2016) 460–473.
- [6] L. Zhou, D. Wu, J. Chen, et al., Cross-modal collaborative communications, *IEEE Wirel. Commun.* 27 (2) (2020) 112–117.
- [7] X. Wei, L. Zhou, AI-enabled cross-modal communications, *IEEE Wirel. Commun.* 28 (4) (2021) 182–189.
- [8] G.J. Sullivan, J.-R. Ohm, W.-J. Han, et al., Overview of the high efficiency video coding (HEVC) standard, *IEEE Trans. Circ. Syst. Vid.* 22 (12) (2012) 1649–1668.
- [9] J.M. Soyce, Y. Ye, J. Chen, et al., Overview of SHVC: Scalable extensions of the high efficiency video coding standard, *IEEE Trans. Circ. Syst. Vid.* 26 (1) (2016) 20–34.
- [10] B. Bross, Y. Wang, Y. Ye, et al., Overview of the versatile video coding (VVC) standard and its applications, *IEEE Trans. Circ. Syst. Vid.* 31 (10) (2021) 3736–3764.
- [11] D.Z. Rodríguez, R.L. Rosa, E.C. Alfaia, et al., Video quality metric for streaming service using DASH standard, *IEEE Trans. Broadcast.* 62 (3) (2016) 628–639.
- [12] L. Yu, T. Tillo, J. Xiao, QoE-driven dynamic adaptive video streaming strategy with future information, *IEEE Trans. Circ. Syst. Vid.* 63 (3) (2017) 523–534.
- [13] M. Gdaleta, F. Chiariotti, M. Rossi, et al., D-DASH: A deep q-learning framework for DASH video streaming, *IEEE Trans. Cogn. Commun. Netw.* 3 (4) (2017) 703–718.
- [14] L. Zhou, D. Wu, J. Chen, et al., Greening the smart cities: Energy-efficient massive content delivery via d2d communications, *IEEE Trans. Ind. Inform.* 14 (4) (2018) 1626–1634.
- [15] R. Xiong, J. Zhang, F. Wu, et al., Power distortion optimization for uncoded linear transformed transmission of images and videos, *IEEE Trans. Image Process.* 26 (1) (2017) 222–236.
- [16] X. Wei, L. Zhou, *Multimedia QoE Evaluation*, Springer, 2019.
- [17] O. Holland, E. Steinbach, R.V. Prasad, et al., The IEEE 1918.1 “tactile internet” standards working group and its standards, *Proc. IEEE* 107 (2) (2019) 256–279.
- [18] E. Steinbach, S. Hirche, J. Kammerl, et al., Haptic data compression and communication, *IEEE Signal Proc. Mag.* 28 (1) (2011) 87–96.
- [19] N. Sakr, N.D. Georganas, J. Zhao, Human perception-based data reduction for haptic communication in six-DoF telepresence systems, *IEEE Trans. Instrum. Meas.* 60 (11) (2011) 3534–3546.
- [20] R. Hassen, B. Gülecüyüz, E. Steinbach, PVC-SLP: Perceptual vibrotactile-signal compression based on sparse linear prediction, *IEEE Trans. Multimedia* 23 (2020) 4455–4468.
- [21] E. Steinbach, M. Strese, M. Eid, et al., Haptic codecs for the tactile internet, *Proc. IEEE* 107 (2) (2019) 447–470.
- [22] R. Chaudhari, C. Schuwerk, M. Danaei, et al., Perceptual and bitrate-scalable coding of haptic surface texture signals, *IEEE J. Sel. Top. Signal Process.* 9 (3) (2015) 462–473.
- [23] J. Xu, K. Ota, M. Dong, Energy efficient hybrid edge caching scheme for tactile internet in 5G, *IEEE Trans. Green Commun. Netw.* 3 (2) (2019) 483–493.
- [24] Y. Xiao, M. Krunz, Distributed optimization for energy-efficient fog computing in the tactile internet, *IEEE J. Sel. Areas Commun.* 36 (11) (2018) 2390–2400.
- [25] X. Wei, Q. Duan, L. Zhou, A QoE-driven tactile internet architecture for smart city, *IEEE Netw.* 34 (1) (2020) 130–136.
- [26] V. Fanibhare, N.I. Sarkar, A. Al-Anbuky, Toward a fog-based traffic flow framework for tactile internet, *IEEE Internet Things J.* 9 (3) (2022) 10718–10731.

- [27] X. Li, Z. Yuan, J. Zhao, et al., Edge-learning-enabled realistic touch and stable communication for remote haptic display, *IEEE Netw.* 35 (1) (2021) 141–147.
- [28] M. Mukherjee, M. Guo, J. Lloret, et al., Leveraging intelligent computation offloading with fog/edge computing for tactile internet: Advantages and limitations, *IEEE Netw.* 34 (5) (2020) 322–329.
- [29] Z. Hou, C. She, Y. Li, et al., Intelligent communications for tactile internet in 6G: Requirements, technologies, and challenges, *IEEE Commun. Mag.* 59 (12) (2021) 82–88.
- [30] R.W.L. Coutinho, A. Boukerche, Design of edge computing for 5G-enabled tactile internet-based industrial applications, *IEEE Commun. Mag.* 60 (1) (2022) 60–66.
- [31] Z. Hou, C. She, Y. Li, et al., Burstiness-aware bandwidth reservation for ultra-reliable and low-latency communications in tactile internet, *IEEE J. Sel. Areas Commun.* 36 (11) (2018) 2401–2410.
- [32] A. Aijaz, Hap-slicer: A radio resource slicing framework for 5G networks with haptic communications, *IEEE Syst. J.* 12 (3) (2018) 2285–2296.
- [33] M. Tang, L. Gao, J. Huang, Enabling edge cooperation in tactile internet via 3C resource sharing, *IEEE J. Sel. Areas Commun.* 36 (11) (2018) 2444–2454.
- [34] T. Fang, D. Wu, J. Chen, et al., Joint distributed cache and power control in haptic communications: A potential game approach, *IEEE Internet Things J.* 8 (18) (2021) 14418–14430.
- [35] Z. Xiang, F. Gabriel, E. Urbano, et al., Reducing latency in virtual machines: Enabling tactile internet for human-machine co-working, *IEEE J. Sel. Areas Commun.* 37 (5) (2019) 1098–1116.
- [36] A.S. Shafiq, B. Lorenzo, S. Glisic, et al., A framework for dynamic network architecture and topology optimization, *IEEE/ACM Trans. Netw.* 24 (2) (2016) 717–730.
- [37] G. Mountaser, T. Mahmoodi, O. Simeone, Reliable and low-latency fronthaul for tactile internet applications, *IEEE J. Sel. Areas Commun.* 36 (11) (2018) 2455–2463.
- [38] H. Chung, H.H. Lee, K.O. Kim, et al., TDM-PON-based optical access network for tactile internet, 5G, and beyond, *IEEE Netw.* 36 (2) (2022) 76–81.
- [39] A. Aijaz, M. Dohler, A.H. Aghvami, et al., Realizing the tactile internet: Haptic communications over next generation 5G cellular networks, *IEEE Wirel. Commun.* 24 (2) (2017) 82–89.
- [40] P. Olsson, F. Nysjö, I.B. Carlbom, et al., Comparison of walking and traveling-wave piezoelectric motors as actuators in kinesthetic haptic devices, *IEEE Trans. Haptics* 9 (3) (2016) 427–431.
- [41] S. Sundaram, P. Kellnhofer, Y. Li, et al., Learning the signatures of the human grasp using a scalable tactile glove, *Nature* 569 (7758) (2019) 698.
- [42] M. Strese, R. Hassen, A. Noll, et al., A tactile computer mouse for the display of surface material properties, *IEEE Trans. Haptics* 12 (1) (2019) 427–431.
- [43] K. Yao, J. Zhou, Q. Huang, et al., Encoding of tactile information in hand via skin-integrated wireless haptic interface, *Nature Mach. Intell.* 4 (10) (2022) 893.
- [44] Z. Sun, M. Zhu, X. Shan, et al., Augmented tactile-perception and haptic-feedback rings as human-machine interfaces aiming for immersive interactions, *Nature Commun.* 13 (1) (2022) 5224.
- [45] R. Hassen, E. Steinbach, HSSIM: An objective haptic quality assessment measure for force-feedback signals, in: *Proceedings of International Conference on Quality of Multimedia Experience*, 2018, pp. 1–6.
- [46] X. Liu, M. Dohler, Y. Deng, Vibrotactile quality assessment: Hybrid metric design based on SNR and SSIM, *IEEE Trans. Multimedia* 22 (4) (2020) 921–933.
- [47] R. Hassen, E. Steinbach, Subjective evaluation of the spectral temporal similarity (ST-SIM) measure for vibrotactile quality assessment, *IEEE Trans. Haptics* 13 (1) (2020) 25–31.
- [48] M. Eid, J. Cha, A.E. Saddik, Admux: An adaptive multiplexer for haptic-audio-visual data communication, *IEEE Trans. Instrum. Meas.* 60 (1) (2020) 21–31.
- [49] B. Cizmeçi, X. Xu, R. Chaudhari, et al., A multiplexing scheme for multimodal teleoperation, *ACM Trans. Multim. Comput.* 13 (2) (2017) 21.
- [50] L. Zhou, Z. Yuan, Cross-modal coding, *J. Nanjing Univ. Post. Telecommun.* 40 (5) (2020) 95–100.
- [51] Z. Yuan, B. Kang, X. Wei, et al., Exploring the benefits of cross-modal coding, *IEEE Trans. Circ. Syst. Vid.* 32 (12) (2022) 8781–8794.
- [52] L. Zhou, D. Wu, X. Wei, et al., Cross-modal stream scheduling for ehealth, *IEEE J. Sel. Areas Commun.* 39 (2) (2021) 426–437.
- [53] Y. Suo, Y. Chen, Y. Gao, et al., Dynamic transmission mode selection for multi-modal services, *IEEE Commun. Lett.* 27 (3) (2023) 911–915.
- [54] L. Yang, D. Wu, L. Zhou, Heterogeneous stream scheduling for cross-modal transmission, *IEEE Trans. Commun.* 69 (9) (2021) 6037–6049.
- [55] D. Wu, L. Zhou, Cross-modal stream transmission: Architecture, strategy, and technology, *IEEE Wirel. Commun.* (2023), doi:10.1109/MWC.013.2200293.
- [56] X. Wei, M. Zhang, L. Zhou, Cross-modal transmission strategy, *IEEE Trans. Circ. Syst. Vid.* 32 (6) (2022) 3991–4003.
- [57] X. Wei, J. Zhao, L. Zhou, et al., Broad reinforcement learning for fast autonomous IoT, *IEEE Internet. Things J.* 7 (8) (2020) 7010–7020.
- [58] Y. Gao, X. Wei, B. Kan, et al., Edge intelligence empowered cross-modal streaming transmission, *IEEE Netw.* 35 (2) (2021) 236–243.
- [59] X. Wei, Y. Shi, L. Zhou, Haptic signal reconstruction for cross-modal communications, *IEEE Trans. Multimedia* 24 (2022) 4514–4525.
- [60] A. Li, Y. Chen, S. Ni, et al., Haptic signal reconstruction in ehealth internet of things, *IEEE Internet. Things J.* 9 (18) (2022) 17047–17057.
- [61] A. Li, J. Chen, X. Wei, et al., 6G-oriented cross-modal signal reconstruction technology, *J. Commun.* 43 (6) (2022) 28–40.
- [62] X. Wei, Y. Yao, H. Wang, et al., Perception-aware cross-modal signal reconstruction: From audio-haptic to visual, *IEEE Trans. Multimedia* (2022), doi:10.1109/TMM.2022.3194309.
- [63] J. Xu, X. Wei, L. Zhou, Information recovery technology for cross-modal communications, *Acta Electron. Sin.* 50 (7) (2022) 1631–1642.
- [64] Y. Gao, S. Ni, D. Wu, et al., Edge-based cross-modal communications for remote healthcare, *IEEE J. Sel. Areas Commun.* 40 (11) (2022) 3139–3151.
- [65] K. Niu, J. Dai, S. Yao, et al., A paradigm shift toward semantic communications, *IEEE Commun. Mag.* 60 (11) (2022) 113–119.
- [66] G. Shi, Y. Xiao, Y. Li, et al., From semantic communication to semantic-aware networking: Model, architecture, and open problems, *IEEE Commun. Mag.* 59 (8) (2021) 44–50.
- [67] X. Luo, H.-H. Chen, Q. Guo, Semantic communications: Overview, open issues, and future research directions, *IEEE Wirel. Commun.* 29 (1) (2022) 210–219.
- [68] H. Xie, Z. Qin, A lite distributed semantic communication system for internet of things, *IEEE J. Sel. Areas Commun.* 39 (1) (2021) 142–153.
- [69] Z. Weng, Z. Qin, Semantic communication systems for speech transmission, *IEEE J. Sel. Areas Commun.* 39 (8) (2021) 2434–2444.
- [70] M. Jankowski, D. Gündüz, K. Mikołajczyk, Wireless image retrieval at the edge, *IEEE J. Sel. Areas Commun.* 39 (1) (2021) 89–100.
- [71] T.-Y. Tung, D. Gündüz, DeepWiVe: Deep-learning-aided wireless video transmission, *IEEE J. Sel. Areas Commun.* 40 (9) (2022) 2570–2583.
- [72] D. Gündüz, Z. Qin, I.E. Aguerri, et al., Beyond transmitting bits: Context, semantics, and task-oriented communications, *IEEE J. Sel. Area. Commun.* 41 (1) (2023) 5–41.
- [73] A. Li, X. Wei, D. Wu, et al., Cross-modal semantic communications, *IEEE Wirel. Commun.* 29 (6) (2022) 144–151.
- [74] Y. Gao, X. Wei, J. Chen, et al., Towards immersive experience: Evaluation for interactive network services, *IEEE Netw.* 36 (1) (2022) 144–150.
- [75] Y. Gao, D. Wu, L. Zhou, How to improve immersive experience? *IEEE Trans. Multimedia* 9 (17) (2022), doi:10.1109/TMM.2022.3199666.
- [76] L. Zhao, D. Wu, L. Zhou, Quality-of-decision-driven machine-type communication, *IEEE Internet. Things J.* 9 (17) (2022) 16631–16642.
- [77] Y. Gao, J. Liao, X. Wei, et al., Quality-aware massive content delivery in digital twin-enabled edge networks, *China Commun.* 20 (2) (2023) 1–13.

Author profile

Xin Wei received his PhD degree major at information and communication engineering from Southeast University, Nanjing, China, in 2009. He is currently a professor in Nanjing University of Posts and Telecommunications, China. His research interests are in the areas of multimedia communications, educational technology.

Liang Zhou (BRID: 08351.00.91229) received his PhD degree major at electronic engineering both from École Normale Supérieure (E.N.S.), Cachan, France and Shanghai Jiao Tong University, Shanghai, China, in 2009. He is currently a professor in Nanjing University of Posts and Telecommunications, China. His research interests are in the area of multimedia communications and computing.