IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 24, 2021, accepted December 23, 2021, date of publication January 4, 2022, date of current version January 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3140175

A Metaverse: Taxonomy, Components, Applications, and Open Challenges

SANG-MIN PARK[1] AND YOUNG-GAB KIM[2], (Member, IEEE)
[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, South Korea
[2]Department of Computer and Information Security, and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, South Korea

Corresponding author: Young-Gab Kim (alwaysgabi@sejong.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean Government (MSIT) under Grant 2021R1A2C2012635.

ABSTRACT Unlike previous studies on the Metaverse based on Second Life, the current Metaverse is based on the social value of Generation Z that online and offline selves are not different. With the technological development of deep learning-based high-precision recognition models and natural generation models, Metaverse is being strengthened with various factors, from mobile-based always-on access to connectivity with reality using virtual currency. The integration of enhanced social activities and neural-net methods requires a new definition of Metaverse suitable for the present, different from the previous Metaverse. This paper divides the concepts and essential techniques necessary for realizing the Metaverse into three components (i.e., hardware, software, and contents) and three approaches (i.e., user interaction, implementation, and application) rather than marketing or hardware approach to conduct a comprehensive analysis. Furthermore, we describe essential methods based on three components and techniques to Metaverse's representative Ready Player One, Roblox, and Facebook research in the domain of films, games, and studies. Finally, we summarize the limitations and directions for implementing the immersive Metaverse as social influences, constraints, and open challenges.

INDEX TERMS Artificial intelligence, metaverse, cyber world, avatar, extended reality.

I. INTRODUCTION

Metaverse is expanding rapidly, as seen in Geppetto serving 200 million subscribers and Animal Crossing running an election campaign in a virtual space. In particular, Roblox's monthly active users (MAU) is 150 million, which is used by 2/3 of children aged 9-12 in the US, and 1/3 of them are under 16 [1]–[3]. Early studies for the Metaverse focus on Second Life in 2006 [4]–[6]. However, the current Metaverse is based on the social values of Generation Z that online ego is no different from offline ones [7]. Therefore, since the proportion of social activities and contents grows, it differs from the previous Metaverse, and a new definition is needed for the present.

The novel Metaverse differs from the earlier Metaverse in three ways. First, the rapid development of deep learning dramatically improves the accuracy of vision and language recognition, and the development of generative models enables a more immersive environment and natural movement. The processing time and complexity were reduced using multimodal models as E2E (end-to-end) solutions with a multimodal pre-trained model. Second, Metaverse previously served based on PC access and had low consistency due to time and space constraints, but now it is possible to easily access the Metaverse anytime, anywhere due to the mobile devices that can connect to the Internet at all times. There are 50 million games in Roblox and the accumulated monthly usage time is 3 billion hours. People consums more time than social network services (e.g., TikTok, YouTube). It has a virtuous cycle ecosystem in which the inflow and income of producers increase as users and usage time increase while serving various contents, and thus sales of digital advertisements increase. Lastly, the current Metaverse differs from the previous one because the program coding can be done in the Metaverse world, and it is more bonded to real life with virtual currency. Metaverse expands with various social meanings (e.g., fashion, event, game, education, and office) based on immersive interaction.

The associate editor coordinating the review of this manuscript and approving it for publication was Sudhakar Babu Thanikanti.

Cryptocurrencies (e.g., Dime) serve as an economic bridge between the Metaverse and the real world, giving people deeper social meaning.

The Metaverse differs from augmented reality (AR) and virtual reality (VR) in three ways. First, while VR-related studies focus on a physical approach and rendering, Metaverse has a strong aspect as a service with more sustainable content and social meaning. Second, the Metaverse does not necessarily use AR and VR technologies. Even if the platform does not support VR and AR, it can be a Metaverse application. Lastly, the Metaverse has a scalable environment that can accommodate many people is essential to reinforce social meaning. The large-scale Metaverse implementation required three components: (i) hardware improvements (e.g., GPU memory, 5G); (ii) the development of the recognition and expression model that leverages the parallelism of the hardware; and (iii) the availability of content that people immerse in and participate in.

Despite the considerable research relating to Metaverse, primarily focus on social meaning, and little attention has focused on technologies for the Metaverse. For example, a systematic approach to what concepts and technologies are required to create an environment and content that users can enjoy in Ready Player One is needed. Beyond simply creating a physical, virtual space, it is able to provide an immersive experience with a story through user interaction. This research presents a comprehensive study on the applications and technologies that can give social meaning in a Metaverse hardware, software, and content with three approaches (i.e., user interactions, implementation, and applications).

Firstly, this study analyzed hardware components, software components, and contents into component levels to create an immersive experience in the Metaverse, as shown in Fig. 1. In order to give the user a sense of visual immersion, a lightweight head-mounted display (HMD) and physical auxiliary devices are required to use for a long time with high-resolution images [8]. In terms of software, as the delay increases, dizziness and motion sickness occur due to sensory confusion, so low delay and fast rendering are important. In addition, since Metaverse is conducted based on a wide 360-degree field of view, large-capacity vision data processing and generative recognition of obscured objects are significant issues. There is a technical gap in hardware and software performance compared to users' expectations for the Metaverse. The natural movement of the graphical environment displayed in the Metaverse can give an immersive feeling. However, to provide a sustainable service, it is crucial to have immersive contents that work even with limited hardware and low-resolution software (e.g., Minecraft). For sustainable Metaverse content, there must be a plot that considers various user interactions in the virtual environments. In other words, an approach in the form of a complete story (e.g., a movie and a drama) is needed rather than several dialogues turn. Because user-created content is not produced with a large number of organized teams, methods (e.g., persona generation, cartoon generation) to complement professionalism are needed. In addition, multimodal-based stories based on immersive interaction can be used effectively in the Metaverse to implement such an interactive user scenario.

Secondly, this study analyzed user interactions, implementations, and applications into approach levels to provide a stable experience in the Metaverse, as shown in Fig. 2. Comprehensive recognition and interpretation through multimodal inference are required for an interaction that can effectively utilize the technologically growing hardware and software performance. For example, human-robot interaction and visual-language interaction are similar in that they are egocentric views to be used as element technologies for user interactions. Metaverse environments are divided into service platforms (e.g., Roblox, Minecraft) and configurable environments (e.g., Unity) for implementations. In order for the Metaverse to allow many people to live life in the same space, infrastructure elements (e.g., wide bandwidth network connection, fault management, and security) are also important for implementations. The details of each application and event also play an important role in composing the Metaverse. As applications and events (e.g., simulation, marketing, and education) become more concrete, people's activities will increase, and their playing time gradually increases accordingly [9], [10].

Due to the wide scope of the Metaverse, we lack a clear understanding of how they work, why they need, and what they are even capable of due to their novel component. To tackle these problems deeply, we require interdisciplinary collaboration and research with the psychology and social sciences of the Metaverse.

This study has three main contributions as follows.

- Metaverse taxonomy is proposed by summarized technologies and is used to classify the studies of research institutes. We classified the Metaverse components and major approaches into hardware, software, contents, user interaction, implementations, and applications. For each approach, we have summarized the technologies that have recently become issues and interests.
- We classify Metaverse's representative Ready Player One, Roblox, and Facebook research in films, games, and studies using the method defined above and describe the latest technology and development.
- Finally, problems and directions in implementing an immersive Metaverse are divided into social influences, restrictions, and open challenges.

The remainder of this study is organized as follows. In Section 2, various definitions of Metaverse and avatar are arranged in chronological order. Section 3 describes and proposes three components necessary, and Section 4 describes the high-level approach to give an immersive experience in the Metaverse. In Section 5, we verify how the defined components and approaches are used through case studies on Ready player one, Roblox, and Facebook research. Influences, limitations, and open challenges

are discussed in Section 6 and finally concluded in Section 7.

II. METAVERSE CONCEPTS

This section describes the concepts of the Metaverse, avatar, and extended reality (XR) based on differences of similar concepts. The Metaverse refers to the virtual world in which the avatar acts, and the avatar is the user's alter ego and becomes the active subject in the Metaverse. XR is the medium that connects avatars in Metaverse and users in the real world.

A. METHODOLOGY

In this paper, we partially utilized systematic literature reviews (SLRs) techniques to obtain reliable references [11]. The reclusive procedure for selecting references is: 1) search by combining related keywords 2) extract papers that contain keywords in the title and body 3) remove papers that contain keywords but are not directly related to the Metaverse 4) cluster related papers 5) configure taxonomy. At first, we extract keywords (i.e., Metaverse, Avatar, Extended Reality) for Metaverse concepts that are interpreted in various forms, as shown in Table 1. As depicted in Fig. 3, each paper's Metaverse definitions and characteristics are analyzed in chronological order out of a total of 260 papers, including 130 papers of Elsevier papers and 130 papers of Google scholar based on relevance. Table 1 summarizes the definitions and main viewpoints of 54 papers that specifically describe the Metaverse [4], [7]–[10], [12]–[60].

In Section III, we make component taxonomy (i.e., hardware, software, contents) which is necessary to construct the Metaverse with the same procedure including a total of 15 sub-categories. In a similar way, we construct an approach taxonomy (i.e., interaction, implementation, application) incldluing 16 sub-categories for Metaverse approaches in Section IV. Finally, we choose representative services of the Metaverse and evaluate taxonomy by mapping references. Especially in the case of Facebook, we review papers that are announced on the papers published in Facebook Research from January to June 2021.

Duan *et al.* [7] presented the representative applications in the aspect of infrastructure, interaction, and ecosystem. They also provide a three-layer Metaverse architecture containing ana brief timeline for Metaverse development. Messinger *et al.* [23] introduced a virtual world where thousands of people can interact simultaneously within the same simulated 3D space. It covered the perspectives of business, education, social sciences, technical sciences, and social computing that affect our society as a whole. Müller [33] defined the world as an electronic memory and the Internet as a virtual reality where users log in every day. They focus on safely preserving information, the evaluation of data, and perception. Dionisio *et al.* [43] focused on immersive realism, the ubiquity of access and identity, interoperability, and scalability of Metaverse. Nevelsteen [56] focused on ontology as the relation of the complimentary terms and acronyms. They also introduced the usage of pseudo persistence to categories technologies that only mimic persistence. Since various Metaverse surveys mainly focus on application and social meaning, comprehensive research on Metaverse technology is lacking. In order to compose the Metaverse, it is necessary to investigate a comprehensive view of the latest technology components, approaches, and services. We compare Metaverses defined in 54 other surveys in Table 1 and dealt with HW, SW, and Contents in depth. In particular, we evaluate the proposed taxonomy through three different types of use cases.

B. METAVERSE

Metaverse is a compound word of transcendence meta and universe and refers to a three-dimensional virtual world where avatars engage in political, economic, social, and cultural activities. It is widely used in the sense of a virtual world based on daily life where both the real and the unreal coexist [61]. Metaverse was first used in Neil Stevenson's science fiction novel Snow Crash in 1992 and referred to a world where virtual and reality interact and create value through various social activities [62]. As the scope of the Metaverse is wide and continuously growing, various definitions and similar concepts exist. Lee *et al.* [63] divided life-logging, mirror world, augmented reality, and the virtual world according to whether the implemented space is reality-oriented or virtual-centered, and whether the implemented information is external environment information-centered and individual-centered. In previous studies, Metaverse focused on the composition of the virtual world itself (e.g., game), but recently, it is often expressed as a medium for exchanging interests and social interaction centered on content.

Mirror world (e.g., Google Earth, Microsoft Virtual Earth) refers to extending information into the virtual world by realistically reflecting the real world. Mirror World is originated from a book called Mirror Worlds written by David Gelernter in 1992 [64]. The real space where people live is reproduced in digital form, and additional simulation information is added. In other words, the mirror world replicates the appearance of buildings or objects in the real world but has its own properties and functions. Metaverse, multiverse, digital terraforming, and mirror world are conceptually similar but have slightly different meanings depending on where they are used and share some concepts.

C. AVATAR

An avatar means an alter ego that has descended to the earth, and it started from the concept that a fundamental being (e.g., God) changed its form to human. Previously, the avatar was used as a pre-defined exaggerated form in the virtual world rather than reflecting the real world. However, it gradually changes into an ideal form that projects the outward appearance and reflects the ego. An avatar performs a social role suitable for a job and persona in Metaverse. In particular, costumes and items in Metaverse are used as a

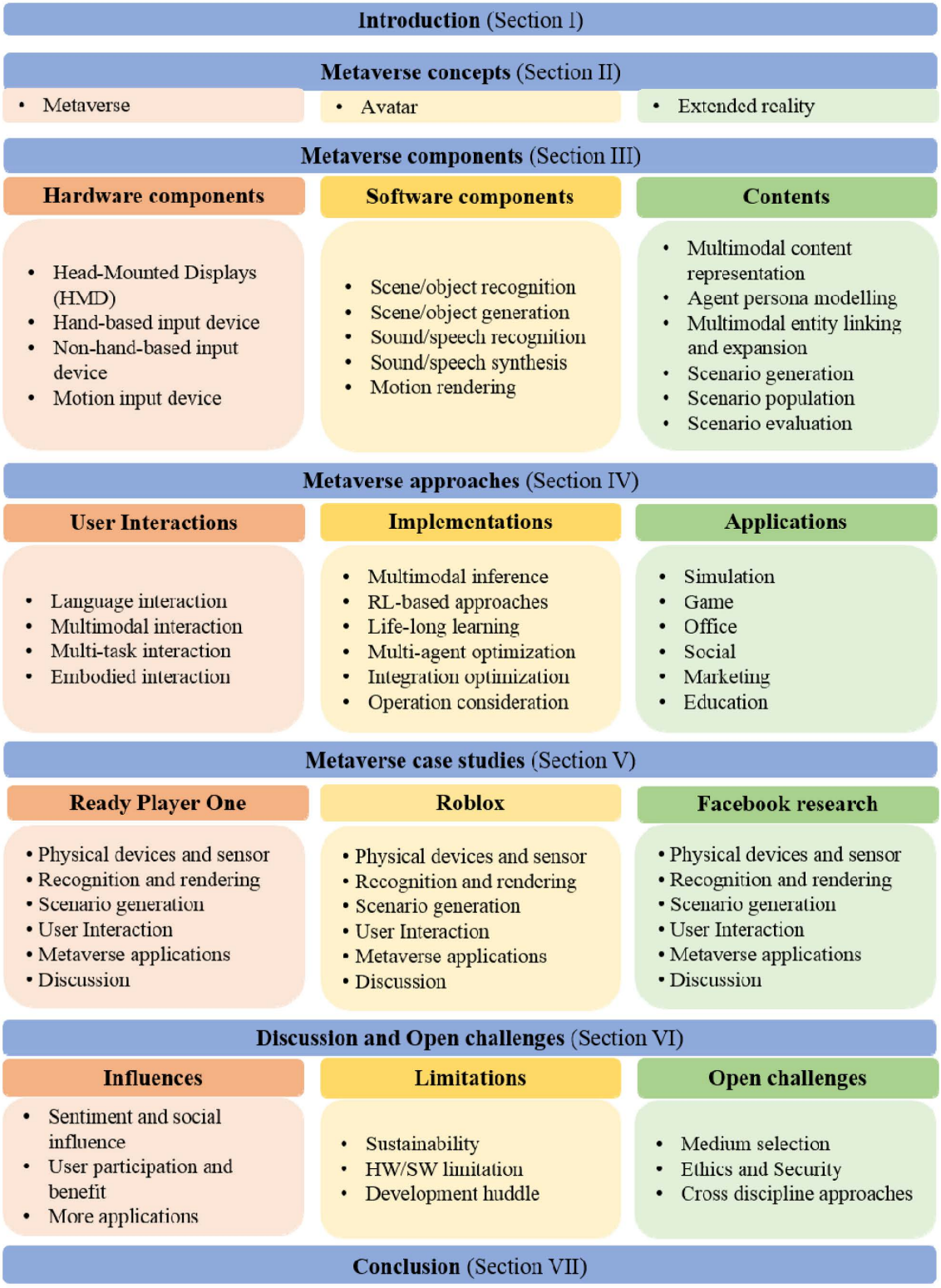

FIGURE 1. Organization of the paper.

medium to express the social meaning of avatars, and various luxury clothing companies are paying attention and selling them. The younger generation considers the social meaning of the virtual world as important as the real world, as they think that their identity in virtual space and reality is the same.

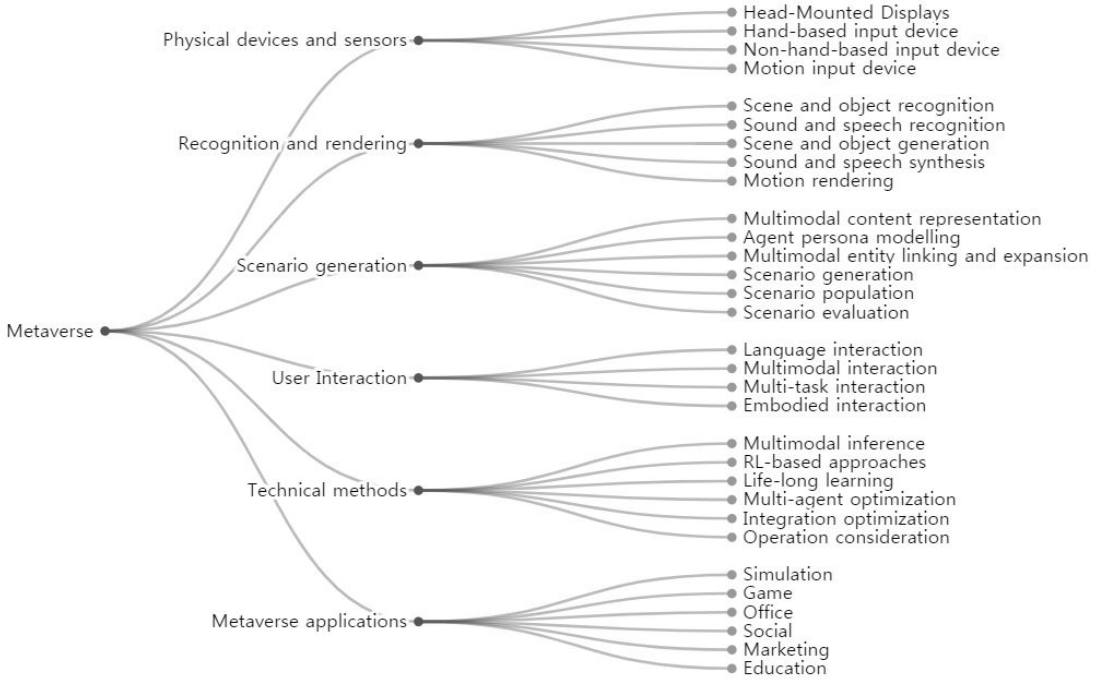

FIGURE 2. Metaverse taxonomy.

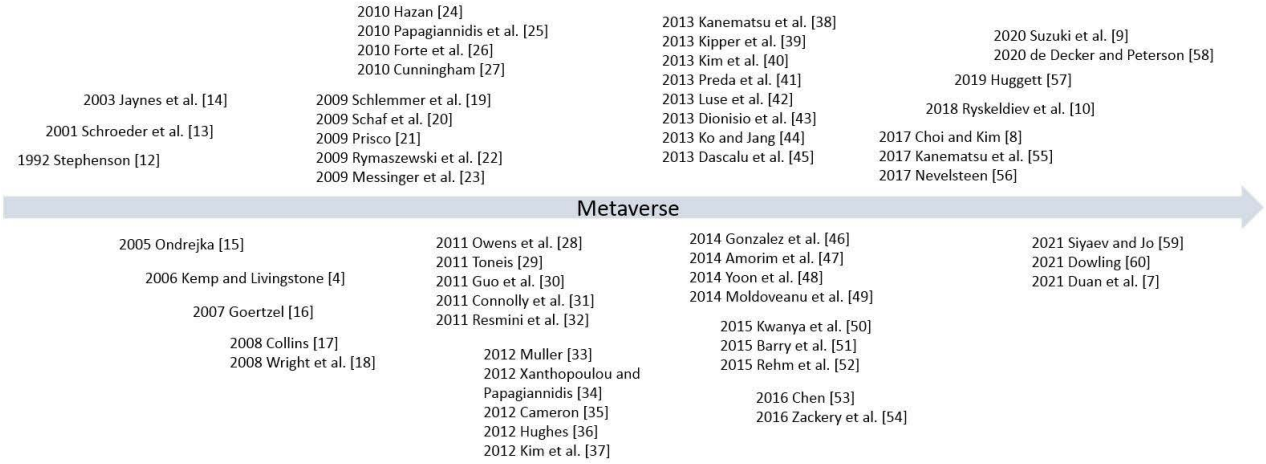

FIGURE 3. Metaverse papers over the time sequence.

Avatar, the subject of the Metaverse, has a similar meaning to the digital twin and digital Me of the virtual world. A digital twin is a virtual model for predicting behavior [65]. Digital twins are used to create real-object-like agents in the virtual world and predict outcomes in advance through simulations of situations that might occur in real life. Initially proposed by General Electronics, the system combines data and information representing contexts and processes of various physical entities to understand past and present operating states. It is used to maintain properties and states throughout the lifecycle of a digital twin and predict what will happen in the future. It can optimize the physical world and is used in various industrial and social issues and manufacturing to improve operational performance and business processes significantly. Digital Me is a symbolic expression of ego in a digital world that is different from the actual self. Conceptually, the digital twin is different in that it objectively interprets the real self, whereas digital Me interprets it subjectively. In terms of application, digital twins are used to solve current problems and simulate future outcomes. Digital Me, on the other hand, is a surrogate self that projects one's self that cannot be done in real life.

D. EXTENDED REALITY (XR)

In terms of technology, XR is related to VR (virtual reality), AR (augmented reality), and MR (mixed reality).

VR used to act as an avatar in a digitally implemented three-dimensional world (e.g., ZEPETO). VR provides an

TABLE 1. Metaverse definition.

Vendor	Definition	Characteristics
Stephenson [12]	A world where humans as avatars interact with each other and with software agents in a three-dimensional space that reflects the real world.	Allowing users to create new entities to have a market value; Describing the tension between the request and ownership between the player and the operator.
Schroeder et al. [13]	A resident virtual world where the geography and physical characteristics of the real world are modeled in a networked digital space where the user is represented as an avatar.	Describing the connection between science fiction and cyberpunk culture comparing 'nonspace.'
Jaynes et al. [14]	An immersive environment using a universal and shared digital media network that removes the barriers of time and space by deceiving users' visual senses	Visually immersive, self-organizing and monitoring, interactive, collaborative capabilities
Ondrejka [15]	The technical challenges of making something close to the complexity and realism depicted in Snow Crash	Potential to open large markets for capital and wealth by empowering users to their creations with dynamic complexity and the right to create content
Kemp and Livingstone [4]	Access online systems as exclusive clients and interact with content and other residents	Links to external web pages and Internet resources, tools for constructing 3D objects, scripting for interactive content
Goertzel [16]	An increasingly intelligent world where AGIs are integrated into interacting human social networks	Artificial intelligence agents are an important part of Metaverse
Collins [17]	From business to entertainment, an interactive network with continuous, immersive 3D virtual environments accessible	Convergence of a virtual, augmented physical reality with a physically persistent virtual space
Wright et al. [18]	Extensive 3D network virtual world that can support many people at the same time for social interaction	Social interaction and collaboration, the interaction between real people and virtual environments and agents and virtual environments, including avatars
Schlemmer et al. [19]	Extension of the parallel space of the physical world within the virtual Internet space into cyberspace	Experience immersion through telepresence as an avatar; The technological incarnation of the old daydream in which parallel worlds, collective memory, images, myths, and symbols chase humans.
Schaf et al. [20]	A world of enhancing the feeling of being in a classroom rather than being an incorporeal observer in a 2D virtual environment.	Using state-of-the-art technology to support collaboration, creativity, and sharing over the web
Prisco [21]	A complete video-realistic medium based on virtual reality allows immersive interaction between participants.	Sustainable and accelerated using realistic consumer VR technology
Rymaszewski et al. [22]	An environment where you can create your personality, quickly visit different places, explore expansive buildings, and shop your way.	
Messinger et al. [23]	A virtual world where thousands of people can interact simultaneously within the same simulated 3D space.	Business, education, social science, technical science, and social computing impact our society as a whole
Hazan [24]	A place where users log in all the time to interact with others in play, commerce, creativity, and exploration.	Fringe for the escapist a persistent world beyond the illusion
Papagiannidis et al. [25]	A continuous, continuous world designed to give users control over almost every aspect of the world by creating the objects they want	A vibrant, dynamic world with creative, self-expression, and exciting content that supports different types of applications based on themes.
Forte et al. [26]	A virtual place where an individual's cyber community can share social interactions without the constraints of the physical world.	Addressing scalability, access levels, inter-agent communication, social rules and conventions shared by users, and economic activity; A virtual art museum of the Roman city of Interamnia as a cultural metaverse
Cunningham [27]	A compound word of meta and universe, meaning beyond, a temporal-spatial aspect where the real world and the virtual world are mixed.	Computing everywhere means information everywhere, and all things are digitized through ubiquitous computing technology.
Owens et al. [28]	An immersive three-dimensional virtual world in which people interact with each other and their environment, using real-world metaphors but without physical limitations.	
Tonéis [29]	A world that reconstructs the meaning of the living world with the experience	Consequences of actions, decisions, or choices with aesthetic experience reflect temporally apparent consequences; Consequences construct thinking into ontological aspects in the form of organizing and building knowledge.
Guo et al. [30]	A computer simulation that allows avatars to interconnect and communicate in relatively life-like environments	
Connolly et al. [31]	Continuous online 3D world	A downloadable client program to access the system and interact with content and other residents through customizable avatars
Resmini et al. [32]	One of the variants of the Matrix movie with some good swordsmanship or some zero-gravity kung fu.	Information leaks to the Internet and the real world via mobile phones, pads, public real-time displays, consumer electronics, and connected device
Müller [33]	A world like electronic memory and the Internet as a virtual reality where users log in every day.	An infrastructure for electronic memory in the context of the next-generation Internet; Cannot be in two places at once; Can only move at a limited speed, in restricted areas.
Xanthopoulou and Papagiannidis [34]	A three-dimensional extension of the traditional electronic space that typically hosts massively multiplayer online role-playing games (MMORPGs).	The avatar is the player's virtual persona.

experience as if you were in a specific place without physical limitations, helping you learn about the ideas you can get from experiencing different places. While VR is a technology that allows a new reality to compete based on 360-degree

TABLE 1. *(Continued.)* **Metaverse definition.**

Cameron [35]	Utopian and dystopian futures, where people live more in virtual worlds than in reality	
Hughes [36]	An asynchronous environment that users connect to and an avatar-connected world that is a proxy for a digitally represented human being.	Moving in the environment gives the user a different view of the virtual world, which is visible to other people.
Kim et al. [37]	A collective online space created by combining some physical reality enhanced by a 3D virtual world and a physically permanent virtual space.	Includes all virtual worlds, augmented reality, and the Internet
Kanematsu et al. [38]	A 3D virtual space where the avatar is activated on behalf of the user.	Second Life as an example
Kipper et al. [39]	Cyberspace where everyone is interconnected, similar to the Internet accessed through a medium called virtual reality.	Includes simulations, WWW, different types of interfaces, collaborative environments, and other kinds of worlds.
Kim et al. [40]	The virtual world which connects physical devices (e.g., biosensors)	Use cases of physical exercise
Preda et al. [41]	Collective online shared space	Convergence of virtual, augmented reality, and physically permanent virtual space, including the sum of all generated VW, AR, and Internet
Luse et al. [42]	Virtual world technology that allows you to live your virtual life online	
Dionisio et al. [43]	An integrated network of 3D virtual worlds in an independent virtual world or an attractive alternative realm for human sociocultural interaction.	Features realism, ubiquity, interoperability, and scalability
Ko and Jang [44]	An online virtual community that allows the use of simulations and objects to interact with other users through avatars.	Interactivity, physical persistence, online chatting, entertainment, and educational goals.
Dascalu et al. [45]	New environments and visualizations where physical and digital objects co-exist and interact in real-time	Suitable for modern educational application, raising the efficiency of the learning process
González et al. [46]	Instantiation of a 3D virtual space where people interact with each other via avatars and clients.	Transforming education, learning, virtual project management, and conversation; Control the virtual world with the actions of your avatar, providing reality without the physical limitations of the real world
Amorim et al. [47]	An immersive environment that can simulate real-world features (e.g., sound and gravity)	
Yoon et al. [48]	An immersive world of information where anything you can imagine today is connected to the Internet and intensely stimulates the senses.	Creating and disseminating information, seamlessly merging the virtual and physical worlds; Using AI and feedback systems to enhance human-machine interactions
Moldoveanu et al. [49]	Open 3D platform, consisting in a collection of customized 3D world	Providing 3D visual interface, which provides not only remote access to administrative and education services but also provides their feeling with new interaction and communication
Kwanya et al. [50]	Online shared space created by the convergence	Providing an architecture that enables interoperable multimedia and multi-mode communication
Barry et al. [51]	A virtual 3D world where the avatar does everything for you.	
Rehm et al. [52]	Virtually augmented physical reality and physically persistent virtual space	Taking into account technical, social, legal, economic, and other aspects and factors; A vehicle for change in cyber-physical evolution at various levels
Chen [53]	Immersive environments that reflect the real world and are co-created by residents using their imaginations	
Zackery et al. [54]	A world that can exist in different temporally, politically, and culturally different forms through human-machine interactions enables the game's agents to solve present problems, redefine the past, and invent the future.	Interacting with the environment and users engage through games and socializing without thinking about existence; A virtual debate community of value-driven proxy seekers who communicate without boundaries between human and non-human elements.
Choi and Kim [8]	A space created by the fusion of virtual reality and augmented reality as a compound word of abstract concepts meta and universe	Four key elements: augmented reality, virtual world, lifelogging, and mirror world
Kanematsu et al. [55]	Created world with four different factors: realism, ubiquity, interoperability, and extensibility.	Describing the technical challenges, economic and political barriers of real-world modeling objects in the virtual world.
Nevelsteen [56]	An interactive human-computer mediated simulation of an artificial environment as a permanent, synthetic, 3D, non-game-centric space that separates games and social spaces.	Internet-like, mixed reality into a virtual world (video conferencing, live web cameras in cities, remote operations, projecting buildings from networks)
Ryskeldiev et al. [10]	A constantly updated world of mixed reality spaces mapped to different geospatial locations	Archiving, recycling, and sharing virtual spaces among various mixed reality applications; Reducing the computational cost of mobile mixed reality applications and expanding interactive space

images, AR is a method of superimposing virtual objects on real space from a first-person perspective (e.g., Pokemon Go). AR overlays computer-generated images, sounds, 3D models, videos, graphics, animated sequences, games, and

TABLE 1. *(Continued.)* **Metaverse definition.**

Huggett [57]	A world where virtual worlds combine immersive VR with physical actors, objects, interfaces, and networks in a future form of the Internet; A social virtual world that parallels and replaces the real world	Including immersive realism, ubiquity (university, availability, accessibility, interoperability), and scalability (number of users, scene complexity, and interaction).
Suzuki et al. [9]	A world of dimensions in which the avatar acts on behalf of the user in the real world.	A virtual world composed of computer graphics can be accessed by users with appropriate personal computers and special applications.
De Decker and Peterson [58]	The conceptual environment in which users of networked computers interact	The convergence of virtual augmented physical reality and physically permanent virtual space
Siyaev and Jo [59]	Stunning mixed reality digital space inside the physical world, interacting in millions of 3D virtual experiences	With the requirement of physical distancing, an online digital environment allows people to share an increasingly diverse human experience.
Dowling et al. [60]	A next-generation virtual world built on blockchain	To implement the Metaverse in blockchain games, monetize the supply and demand markets rather than competition between players.
Duan et al [7]	The world that the users, as avatars, can interact with each other and software applications in a 3D virtual space	Macro point of view, infrastructure, interaction, and ecosystem; a blockchain-driven metaverse prototype

GPS information into real-world environments [66], [67]. Visually search for objects and adjust interfaces by overlaying visually immersive content in the real world. In particular, it has the advantage of clearly providing information and visualizing controllable devices without an additional screen.

MR, which integrates these two concepts, is a mixed reality technology that integrates VR and AR. MR is the concept of creating virtual objects that allow users to interact with the 3D environment in the immersion of the virtual environment of VR and the overlay of virtual content in AR. AR provides a more realistic solution because the hardware is relatively simple, like glasses, and reflects reality well, but it is suitable for short content [68]. On the other hand, VR covers the entire field of view, has an immersive feeling, and is suitable for long-term content but entails physical fatigue. In some cases, MR, which uses a mixture of these advantages and disadvantages, is being considered as a solution that can be converted to AR and VR with a single device. XR is an extended reality, which is terms used to include VR, AR, and MR. XR is used for virtual commerce or v-commerce to create computer-mediated indirect experiences [69].

III. METAVERSE COMPONENTS

Metaverse gives patients an immersive experience enough to be used in psychotherapy. People know that myths and novels are not realistic, but they are moved. Similarly, Metaverse is not the real world but can provide a tangible feeling, so services based on immersive user-interactive stories can provide. A representative example of such an approach is a game based on two-way interaction. In order to service the Metaverse like the real world, it is necessary to be able to interact seamlessly and concurrency in an environment with presence. In order to maintain a sustainable Metaverse, economic activity between users based on these interactions must continue. We describe Metaverse into hardware, software, and contents from the component's point of view in this section.

A. HARDWARE COMPONENTS (PHYSICAL DEVICES AND SENSORS)

Hardware in Metaverse not only plays an important role in the immersive experience but also is a technically limiting barrier. In the Metaverse, hardware is quickly enhanced by the effects of technological advancement, but it still needs improvement compared to the experience of the real world. The essential hardware of Metaverse is an HMD that blocks the view to enable immersive participation. For a more effective visual experience, Birnie *et al.* [70] proposed a fovea rendering method that maintains the central part in high resolution similar to human vision. Critical factors for physical devices and sensors are resolution, the field of view size, and latency. Among them, the most important characteristic is latency, which plays an important role in multimodal interactions, so it should be designed considering the threshold for side effects and time gaps.

1) HEAD-MOUNTED DISPLAYS (HMD)

The HMD shows an image through the display and plays the role of playing the sound through the speaker [8]. HMD is a basic input tool of Metaverse and is divided into Non-see-through HMD, Optical-see-through HMD, and video-see-though HMD [71]. In the case of a method that covers the screen, it provides a sense of immersion in a completely virtual world. Optical-see-through (mainly used in AR) is a method of overlaying the virtual world, and high hardware specifications are required in the process of overlaying. To complement this method, video-see-though HMD is used. These HMD issues are the bulky, expensive, and short battery life of the headset. HMD tracks position and orientation according to the movement of the head and delivers the same change of view as in the virtual world by moving the screen. It is more inaccurate than the method of estimating motion by external measurement due to problems with accuracy and delay time, but it is widely used because it can save space and cost.

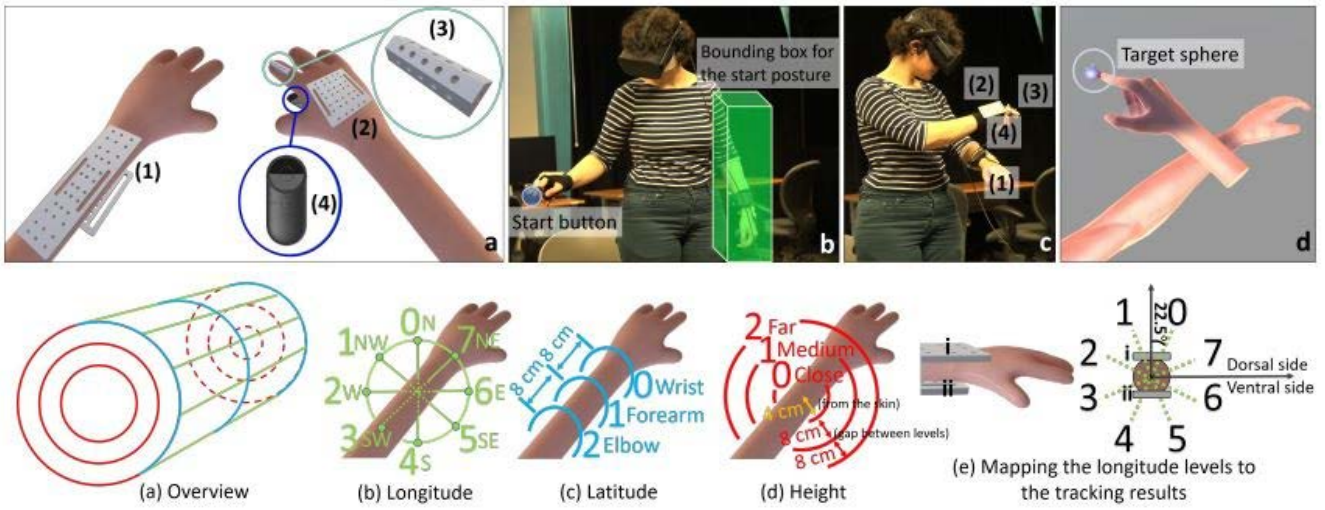

FIGURE 4. The example of circular coordination and area for hand-based input device [72].

2) HAND-BASED INPUT DEVICE

Diverse circular coordination and input area are proposed for hand-based input devices as shown in Fig. 4 [72]. Detailed user data modeling (e.g., mobile phone grip prediction) is required to provide feeling the material with tactile. Haptic has a passive haptic that gives the texture of real objects and an active haptic that creates virtual pressure. Passive haptic is used to help understand the situation while giving presence, and active haptic is used for more effective interaction by adjusting and delivering according to user feedback. Using real props (e.g., physical degree and operational degree) in a virtual environment helps the user experience, while using a robotized interface allows for more diverse interactions [73]. Depending on the device's installation, it is divided into the case of being attached to the hand and the case of being attached to the outside. Beyond making the material feel, it is used in various forms (e.g., inducing muscle tension).

3) NON-HAND-BASED INPUT DEVICE

As auxiliary input means, there are eye-tracking, head tracking, voice input device, and so on [74]. Eye-tracking is a method of changing the viewpoint by predicting eye movement when the user moves their eyes without turning their heads. It is a technology that allows the user to see what kind of object the user is paying attention to. It has the advantage of reducing the load on image processing by generating high-resolution images in the section where the user is focused on a phobia method. The method of overlaying the display on the arm is more stable than the method in the air by repeatedly providing the display at a location predictable by the user [70]. Voice input has an advantage in processing long texts and conversations in a virtual keyboard and an environment where input is limited.

4) MOTION INPUT DEVICE

In order to effectively use the physical sense of space or gravity, body tracking and treadmill are used to provide accurate motion information with auxiliary devices. Motion input devices are also divided into a passive method and an active method. The passive method is a method of delivering a sense to the user with a fixed scenario, and the active method is a method of providing appropriate feedback based on the user's behavior. It is used in various forms to give realism, from a simple way to walking to a 360-degree rotation. There is a risk of injury to the user, so a method of fixing the waist is used with a treadmill.

B. SOFTWARE COMPONENTS (RECOGNITION AND RENDERING)

A cognitive illusion plays an essential role in immersion in the objective reality of the physical space and the subjective reality that users feel. There are two types of cognition: static cognition and dynamic cognition. Static cognition is the proprioceptive senses (e.g., sight, hearing, and touch), while dynamic cognition is sensory balance and body movement [75]. In dynamic cognition, adaptation, attention, and behavior are important features.

According to the object of cognition, it can be divided into the cognition of environment and cognition of an object. In particular, in Metaverse, it is important to reduce the distortion of detection and recognition. Methods for mitigating distortion include changing the shape of the kernel, changing the expression, and increasing the input. Objects of object recognition include faces, poses, gestures, and gazes related to the body. Such object recognition goes through the process of sensing, recording, recognizing, and tracking.

There are two types of stimulation: remote and proximity stimulation. There are bottom-up and top-down approaches to perceiving stimuli. A concept of perception that is distinct from this intuitive sense is also needed. The unconscious

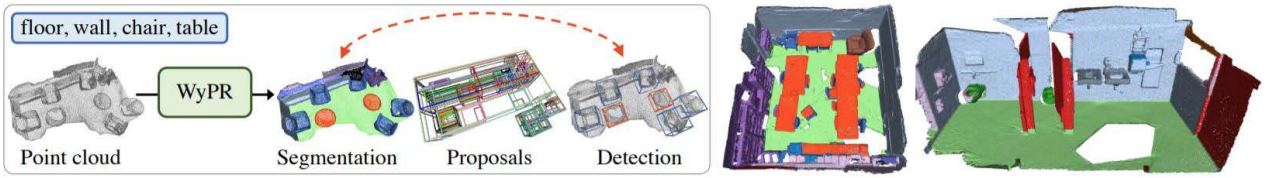

FIGURE 5. Scene rendering for visual language navigation with three-dimension [75].

approach and the conscious approach are classified according to the presence or absence of a difference in movement according to repetitive recognition. There are instinctive, behavioral, contemplative, and emotional processing methods.

The avatar is an important entity in the Metaverse, and the avatar is created, and the action is imitated using animation. Vision-based models estimate human poses, recognize hand gestures and predict gaze. To predict the gaze, iris, facial contour, and 3D gaze prediction are used.

1) SCENE AND OBJECT RECOGNITION

Object recognition is the process of recognizing the size, shape, position, brightness, and colors of objects according to distance. For scene recognition and object recognition, novel methods (e.g., modal alignment, cross-modal attention, point cloud, and scene graph) are used as shown in Fig. 5 [75]. Scene recognition is a good recognition of what state the current scene is and what components and configurations it has. In sub-graph-based scene graph generation, a method of clustering object pairs into graphs by clustering and sharing representations is used [76]. Scene graphs are a good approach to complement the explainable properties that have emerged as limitations of neural network models. Some studies use generative methods and scene graphs to classify bodies in overlapping situations and predict human postures behind walls.

Object recognition is also important along with scene recognition, and we have to pay attention to human-centered scene analysis and non-contact interaction (e.g., gaze, gesture, pose). When many objects are recognized using individual object detection, the number of computations increases in proportion to the number of objects, so an attempt is made to reduce the computational burden by using an abstraction concept. In particular, some studies (e.g., world models and MONET) abstract multiple objects into representations for fast object recognition and efficient training [77].

2) SOUND AND SPEECH RECOGNITION

Recognizing sounds and processing speech help understand surroundings and communicate with other avatars. The conversation is a direct method of communication with other avatars and giving instructions to NPCs in Metaverse. As the Metaverse connection is made in various environments, it is necessary to have a technology that separates the surrounding noise and one's own voice without noise. In addition, the loudness of the sound according to the distance is a variable. For a realistic environment in the Metaverse, voice recognition technology is needed that considers the surrounding environment while adjusting the volume according to the distance.

3) SCENE AND OBJECT GENERATION

The method of generating the environment and objects in Metaverse is divided into the method of depicting by reflecting the real world and the method of creating a new imaginary environment. A realistic way to reflect the real-world environment is to reproduce famous places (e.g., museums, Eiffel Tower) and places familiar to individuals (e.g., home, school) in the real world. Alternatively, it creates a hard-to-reach environment (e.g., underwater, Mars) to provide a surreal experience. People and things are the main objects of object generation. Object generation modules create an avatar and NPC of any desired human shape (e.g., a celebrity, a family member) as an object of conversation. It focuses on facial expressions and natural movements of joints for fluent multimodal conversation. On the other hand, it generates realistic objects that express in detail enough to feel the texture of objects that exist in reality. On the other hand, another type of object is imaginary animals (e.g., unicorns, dragons) and anthropomorphic objects (e.g., talking chairs) that do not exist.

4) SOUND AND SPEECH SYNTHESIS

Sound synthesis is a field that gives the user a sense of immersion, but research is insufficient compared to vision. It creates a sound in the space to give a feeling of presence in the field and to increase the sense of immersion. In particular, a voice suitable for each character is an important means of expressing the character's persona. Tacotron, a speech synthesis, focuses on that users can use prosody to emphasize words or express uncertainty [78]. Prosody is the variation of the speech signal that remains after taking the variation into account (e.g., phonetics and channel effects), which captures meaningful utterances and transfers them by subtractive methods [79].

5) MOTION RENDERING

CNNs and global context encoding are used to capture asymmetric dependencies and context patterns between objects in real-time multi-party 3D motion capture and

pose estimation [80], [81]. The graph reflects the structural characteristics of the body to interpret the action meaning more accurately when the human body is superimposed. Although it is possible to capture the real-time 3D motion of difficult scenes with a single-color camera and isolate human body structures (e.g., shaking hands), it is still limited in capturing close interactions (e.g., hugs).

C. CONTENTS (SCENARIO AND STORY)

Content is the fundamental component that maintains the Metaverse and is used to provide an immersive experience through well-organized stories and user-created events. In content, story reality, immersive experience, and conceptual completeness are important. There are two ways to create content, a paradigm shift method and a method to reuse existing content. The areas that require environment design are scenes, color and lighting, audio, sampling and aliasing, environmental navigation, and real-world content. User motions, characters, and the persona of avatars affect behavioral modeling.

Wang *et al.* [82] introduced studies to process panoramic images and videos in virtual 3D scenes using CNNs and GANs to generate and explore VR content. The generated sentences and images become more natural than before, but sentence patterns with similar meanings are sometimes repeated and evaluated superficially. Also, the longer the sentence, the less the concentration and consistency of the overall content composition. A structured approach (e.g., graph network) is proposed to keep the scenario cohesive and enrich the story's details.

1) MULTIMODAL CONTENT REPRESENTATION

In Metaverse, users create large amounts of multimedia content (e.g., images and videos) as well as text via an avatar. The multimedia data generated in this way expresses the user's thoughts and experiences more than simple dialog. In order to effectively handle multi-modal content, there is an alignment method that converts data into different modal types and a method of expressing data of different modal types by integrating them into one representation [83]. Multimodal content enriches the content by adding information from the data of other modals and supplementing the lack of information of unimodal. By learning these cross-modal features, there is an advantage that intra-mode and inter-modal semantic relationships are utilized.

2) AGENT PERSONA MODELING

In the Metaverse, multi-agents need to have different personas, as if each person has a personality, and multiple agents can interact with in different ways at the same time. It is difficult to give the user a sense of immersion with a character who has a similar conversation every time. Metaverse needs a persona model that expresses various multimodal expressions (e.g., gestures and facial expressions) as well as conversations (e.g., Persona chat). Since spoken language understanding (SLU) uses information on persona without pitch lost in the process of converting a voice signal into text, it grasps a more accurate meaning that is not in the cascaded conversion models.

Although Metaverse users create a lot of user data, entity augmentation and persona generation are important because the data required for learning is relatively large to avoid a cold start problem and the sparsity of various NPC (non-player character) personas. In particular, unevenly user-generated data (e.g., conversation history and personal experience) is biased towards a particular subject until enough data has been gathered.

An entity is a uniquely identified unit (e.g., the name of a famous place and person) and is associated with other entities and relationships. Entity-based expansion is a way to enrich user personas by increasing the number of entities. Methods for increasing the number of entities include generative models, reinforcement learning, joint inference, ontology, and multiple entity extension methods using intermodal [84]–[93]. When tagged resources are scarce, there is a way further to extract entities through co-learning with other model data. Using the pre-trained model is a good way to extend the entity based on the balanced data.

When creating Metaverse NPCs, novel approaches are needed to express persona and emotions that reflect the characteristic of worldview. Therefore, considering the scarcity of each modal, a data population is required to create a balanced persona used in various scenarios. Personas play an important role in giving users a sense of immersion by giving each character a personality in the Metaverse. In particular, it is necessary to implement characters with multi-personas like humans.

When constructing a conversational model, agents responded based on training data for correct answers rather than personality. Because such monotony makes it difficult to maintain long conversations, some researchers proposed to maintain a consistent conversation pattern by introducing the persona concept. For the dialogue system to sustain longer and more human conversations, empathic dialogue systems consider personas [94], [95]. However, dialogue data is insufficient to create large persona representations. Personal Dialog is a large multi-rotational dialogue dataset based on sequential conditional GANs containing different characteristics of different speakers (e.g., age, gender, location, interest tags) [96]. Textual story creation focusing on persona is also proposed [97]. The conditional language model generates various forms (e.g., wiki, horror, humor) of sentences from prefixes without retraining [98].

3) MULTIMODAL ENTITY LINKING AND EXPANSION

When defining characters and entities by transforming the modal of various data, it is necessary to redefine and extend the relationship between the contents through the connection between entities. In expressing the growth process of diverse events and characters in the Metaverse, causality is important in understanding the events and connecting them to the story.

Entity linking is the process of linking related entities based on the similarity between entities and the probability distribution of related contents. Methods for connecting entities by structural learning include link prediction, nonlinear relationships, joint inference, and relationship classification methods [99]–[108]. In particular, the graph model shows how they are related to each other by using the relationship between entities, which are units of meaningful information. The connecting relationship is improved by explicitly modeling the interdependencies between object instances in the scene graph [109]. It also improves the performance of relational inference by encoding global context and geometric layout. Research on graph models and graph convolution networks is increasing to extend the links of stories [110]–[113].

Since the Metaverse contains various worldviews, simply connecting objects is not enough, so the process of expanding and inferring links between entities is required. Inferring information based on given data is an important issue in enriching content. In particular, facts are derived by connecting hierarchically connected things based on causal relationships. Some studies used inference methods include variation inference, various modalities, ontology, emotion, and knowledge [114]–[121].

4) SCENARIO GENERATION

Rather than listing events in the Metaverse, it is important to find hidden relationships based on causal relationships between events and themes and construct a scenario line based on them. Unlike the text-based scenario, the Metaverse is more complex because it has to be configured in multi-modal and embodied environments. Each entity and relationship are used to organize events, and events must be organically combined to form scenario lines. Scenario lines construct the overall structure and serve as an index linking each event. Because it is not just a list of events, the entities and relationships in each event are linked together based on long-term dependencies.

In order to compose a scenario line, it is necessary to connect events composed of entities and their relationships using a graph model. Events are divided into main events and subevents according to their importance in the progress of the scenario. Scenario construction methods include continuous sequences, hierarchical structures, and the attention-based method by focusing on noteworthy content [122]–[132].

When user behavior data in a scenario graph is accumulated over the lifetime of the avatar, it is extended to the concept of life logging. Key scenario topics are extracted with topic modeling and summarized personalized multimodal user data with generative language models. Yu and Riedl [123] introduced a drama manager who personalizes user stories with plots and optimal sequences. Bolanos *et al.* [133] described the visual life logging of storytelling by time slicing, summarizing, and retrieving important information. Li *et al.* [134] proposed StoryGAN, a story-image sequence generation model that sequentially visualizes stories by generating one image sequence for each sentence.

5) SCENARIO POPULATION

Scenarios expand by adding entities and linking the added entities with relation. Scenario lines form a skeleton and expand entities and links to events to create rich stories. Connections between events and other events are formed by relationships and are linked within a scenario. Entity expansion methods include translation embedding, attention, bidirectional inference, and relational inference [135]–[140].

In the process of scenario graph population, modal conversion (e.g., text-to-video and video-to-text conversion) is used for multimodal integration. After pairing sentence nodes with images in a hierarchical approach, it adjusts the length of events through event summarization. The generated multi-mode scenario graph can be used to expand or collapse events. Each event is summarized as representative images with a multimodal language model.

6) SCENARIO EVALUATION

In an event-based extended scenario, as the scenario lengthens, inconsistencies between events occur, it is necessary to verify periodically whether the scenario does not conflict in concept. By instantiating the scenario graph, it hierarchically enlarges and contracts to verify that each event is organically connected and there is no contradiction. Scenario verification is divided with a synthetic method based on grammar and a method directly verifying visualized graphs using human-defined metrics [141]–[146]. Human-defined metrics are divided into a structural approach and a search-based approach. The structural approach evaluates the overall composition in which the scenario is balanced, while the search-based approach looks up specific facts with user queries to ensure that the scenario is well-formed without contradictions.

D. DISCUSSION AND OPEN CHALLENGE

Users can also suffer from simulated motion sickness (i.e., cyber motion sickness) due to the imbalance of visual information obtained from human organs and eyes. There are focusing-displacement collisions and binocular-occlusion collisions, which may have side effects (e.g., blinking). There are other issues (e.g., physical fatigue, headset weight, movement injuries, hygiene issues from prolonged wear) and some side effects (e.g., thin motion sickness, vector motion sickness, eye fatigue, and seizures). In order to reduce side effects, postural stability and physiological measurement methods are used to measure the degree of motion sickness. In addition, adaptive optimization based on the measured values and stabilizing using stable cues have been proposed. Beyond that, there are alternative methods to minimize leading indicators, visual acceleration, and rotation.

Cognitive stability and homeostasis are important for effective service in the Metaverse. Recently, in order to

support a more realistic sense, the scope of Metaverse has been expanded to smell and taste perception. On the other hand, interest in recognizing a complex sense by combining these senses is also growing.

In order to process large amounts of real-time data, fast rendering and data analysis are required immediately. The speed of image processing is essential because the 360-degree field of view is taken into account. Therefore, it is necessary to reduce the delay time through the expected tracking and measurement when rendering the object.

Users are able to decide whether they want to organize their scenarios in a simple, concise summary format or as events in long, complex plots. The depth and length of the scenario are determined using modal and density transforms. Metaverse scenarios use hierarchical and causal relationships to organize events and adjust their length using techniques that adequately summarize the content of a sentence or paragraph. The resolution of the text can be interpreted in terms of the summary, the summary of a scene can consist of a panorama with multiple scenes connected, and it can represent an important timeline among multiple scenes [147]. In terms of scenario construction, studies to find connections between scattered entities include clustering [148], planning-based conditional branching [149], time-dependent index [150], and visual analysis method [151].

The completeness of the content is also an important factor in the Metaverse. For example, in Ready Player One, many interesting characters that dominated popular culture appeared, creating the illusion of going back to that time. However, the story development and probability were weak compared to the splendor of the visual.

IV. METAVERSE APPROCHES

A. USER INTERACTION

Natural interaction is an essential condition for increasing immersion in the Metaverse. It can reproduce the faces of friends and celebrities to enable realistic interactions and to instill the illusion of users with familiar and famous places. Temporary dissociation, concentration, and heightened enjoyment are important factors in the interaction, and emotions of control, curiosity, and intrinsic motivation are used. The target of interaction is mainly human, and hands are an important feature. Input devices are broadly divided into hand-held devices and non-hand input devices. Fidelity, proprioception, and egocentric view are important for interactions on physical devices. Since a 360-degree field of view is used as the receptive field for spatial recognition, a lot of images and distortion corrections are required for video processing efficiency. In order to reduce motion sickness and fatigue, visual and bodily sensory collisions and an alternative sensory method are needed. It also requires multimodal sensory perception that handles speech, gestures, and dialog flows.

1) LANGUAGE INTERACTION

The conversation is a basic approach to deliver user intent via voice recognition. In other words, language is used in various places because it concisely describes complex situations in an implicit sense. It is necessary to create a Metaverse environment in which understanding the situation through language, abstraction, QA, and translation. Miller *et al.* [152] proposed ParlAI, an integrated framework for training and testing conversational models using multitasking training, data collection, human assessment, and online RL. ParlAI performs various tasks in the same interface with dialog datasets (e.g., SQuAD, bAbI task, MCTest, WikiQA, QACNN, QADailyMail, CBT, bAbI dialog, Ubuntu, OpenSubtitles, VQA).

Languages are used in the RL domain as an effective way to define goals and abstract human-comprehensible tasks [153]–[155]. Some agents classify instructions into a single skill level by mimicking human behavior [156]. When the agent is faced with an ambiguous situation, the agent clarifies the instruction intention through a multi-turn conversation with the Oracle [157]. Jiang *et al.* [158] used a language that is flexibly applied to the generalization of various goals, rapid training, and combinations as an abstraction to solve the difficulty of generalized abstraction in hierarchical RL.

AQM (Questioner's Mind) agents ask more consistent questions to maximize information acquisition in task-oriented conversational systems [159]. Knowledge Graph A2C (KG-A2C) is a scalable exploratory method for inferring game states in a template-based workspace using linguistic behaviors and dynamic knowledge graphs [160]. Translation is an essential method in the Metaverse environment where people of various languages are gathered. Domhan *et al.* [161] proposed joint training on a large number of unpaired languages and a small number of language pairs to improve neural machine translation (NMT) performance.

2) MULTIMODAL INTERACTION

Humans facilitate efficient adaptation and reason more abstractly by transferring knowledge across tasks. People communicate not only dialogue but also based on multimodal information (e.g., facial expressions, gestures, and tone of voice). The method of handling each modal is difficult to handle multiple complex emotions, so multimodal interaction is required. In general, multimodal has more information than unimodal and is advantageous for understanding the situation. Text and images in social media posts do not have the same meaning but instead have more complex meanings that intersect semantically [162]–[164]. In particular, multimodal learning is most effective when the meanings of images and text are different.

After the advent of Transformers, studies have been conducted to learn vision and language together and reduce learning from scratch using a pre-trained model. Zhou *et al.* [165] proposed a unified vision language

dictionary training (VLP) model using a shared multi-layer transformer that fine-tunes vision language generation.

3) MULTI-TASK INTERACTION

Since the Metaverse handles many things in the cyber world, a model that handles multiple tasks simultaneously is useful in the aspect of complexity. For such a model, knowledge distillation is used to make a small model that performs many functions and handles other modal types (e.g., Visual QA). Hessel *et al.* [166] argued that multitasking is more complex than single-tasking because the multitasking model balances various tasks in limited expression. It is relatively easy to use for similar tasks but easily overfits when target domain data is scarce and has a different distribution.

E2E methods are also used to perform various tasks effectively. Translatotron [167] translated from voice input to voice output through a sequential process. Compared to the cascaded model, the E2E model has the advantage that most of the inputs can be utilized without data loss in the process. Translatotron interprets a foreign language, including its unique pronunciation and emotional meaning. Also, it has the advantage of responding in a voice form that reflects the prosody of the actual speaker. Qian *et al.* [168] proposed an E2E modeling method SLU for a cloud-based modular dialog system (SDS), showing that it is effective in situations with low ASR accuracy.

4) EMBODIED INTERACTION

The difference between the Metaverse and other general interactions is that the proportion of embodied interactions (e.g., embedded QA and visual language navigation) is relatively high. While the required skills are similar to EQA and VLN, there is a difference in whether the subject is active or passive. While the purpose of VQA is to answer text questions about a given image, EQA (embodied question and answer) performs the task of analyzing sensor information obtained by an agent materialized through active exploration. For example, to answer a question about the color of a car at a distance, the agent actively moves, recognizes, and responds based on prior knowledge of the car's location and path [169]. These EQA tasks have recently been extended in the form of conversations, where agents compensate by querying oracles for insufficient information to perform the task [157].

The factor that differentiates embedded interaction from 2D-based methods is Exophora resolution. Anaphora resolution is the task of analyzing the word in the preceding sentence pointed to by a pronoun [170], [171]. Anaphora and co-reference resolutions are used to infer cross-references in questions and conversations [172], [173]. In short sentences, implied conversations, anaphora resolution is needed to understand the context of the conversation. Recently, such anaphora has been widely used in multi-modal content (e.g., video) and SNS services (e.g., Twitter) beyond simple sentence-based analysis [174], [175]. Exophora resolution maps the meaning of Co-reference resolution and Anaphora resolution used in language to 3D space.

People communicate information in a non-verbal form by pointing to an object instead of language. When a user points to a specific location through a finger, it becomes an intended instruction. In the case of exophora resolution, specific instructions are performed in terms of multimodal interaction, including motion and speech, whereas anaphora simply links meaning between texts. Heinrich *et al.* [176] proposed Embodied Multi-modal Interaction in Language learning (EMIL), a neurocognitive model that reflects in vivo-inspired mechanisms (e.g., an implicit adaptation of time scales).

B. METAVERSE IMPLEMENTATION

The process of Metaverse implementation is divided into a design phase, a model-training phase, an operation phase, and an evaluation phase. The design phase considers goals and concept design, development time and cost, risk estimates, constraints, user scenarios, scope and requirements, and feasibility of implementation and evaluation. In the model-training phase, data analysis, user modeling, scientific methodology, iterative learning, and parameter tuning are performed. The operation phase considers system considerations, simulations, job scheduling, network environments, and prototype demonstrations. The evaluation phase deals with content fidelity, the authenticity of interactions, implementation feasibility, and failover.

This survey covers three types of multimodal inference, RL-based approaches, and lifelong learning for Metaverse training models. In addition, it is necessary to consider multi-agent optimization, integration optimization, and operational considerations from the perspective of Metaverse service operation.

1) MULTIMODAL INFERENCE

Humans do not only interpret the meaning of utterances when communicating with others. When information is given from the cognitive model, it interprets its meaning, combines it with its knowledge, and inferences its intentions. Verbal ambiguity is compensated to determine the speaker's underlying intentions based on direct or indirect representations of the surrounding environment. For example, emotion recognition, the initiator of emotional interaction, uses multimodal fusion to compensate for the lack of context in textual information [177]. Multi-modal models do not always outperform single-mode models, so they should be utilized according to the situation. Zhang *et al.* [178] used late fusion to explicitly examine the impact of each function by considering three types of visual, spatial, and semantic. Liang *et al.* [179] proposed Multimodal Local-Global Ranking Fusion (MLRF), relative sentiment analysis for complex combinations of visual and acoustics. Rather than simply classifying emotions as scalar values, the ranking was performed after measuring the degree of increase or decrease in emotional intensity for partial video segments.

The advantage of the pre-trained model is that it simplifies the task with E2E and does not have to learn from scratch.

Recently, DialogGPT and Vlbert are proposed to implement dialog and visual-language tasks more conveniently. Large-scale pre-trained language models (PLMs) (e.g., Bidirectional Encoder Representations from Transformers (BERT), GPT-3) are used for downstream tasks by applying fine-tuning and few-shot learning [180]. Sun *et al.* [181] propose MobileBERT to compress and accelerate the BERT model. MobileBERT uses the knowledge transfer student model from the teacher BERT large model. Brown *et al.* [182] proposed GPT-3, a 175 billion parameterized autoregressive model that applies several pieces of training runs without any gradient updates and fine-tuning for downstream tasks.

Tan *et al.* [183] proposed contextual mapping of language tokens and associated images with vokenization and multi-modal alignment. It is applied to a relatively small image caption dataset using the generated model.

T5, which integrates text tasks into one model, is proposed to handle translation, question-and-answer (QA), etc. [184]. Video pre-trained models (VPMs) that contain multimodal data of vision and text are effectively used in low-complexity downstream tasks (e.g., VideoBERT [185], ViLBERT [186]). VPMs are used for answering visual questions, common sense reasoning, reference representation, and caption-based image retrieval. BERT-based VPM performs vector quantization of video data and trains bidirectional joint distributions for visual and verbal token sequences.

Some studies give a sense of presence in Metaverse. Domain knowledge understanding provides a more detailed response based on facts. Spoken language understanding deals with the user's tone and emotions. Acoustic signal understanding recognizes and generates sounds from the surrounding environment. Reasoning (e.g., multi-hop reasoning, relational reasoning, and graph reasoning) derived new facts through prior knowledge, background knowledge, and environmental factors given in the current situation.

Multi-hop inference in graph neural networks (GNN) has been used to generate new knowledge about vision and language [187]–[192]. Because graphs, along with KB, act as a repository of knowledge, it is important to effectively utilize encoding, sampling, and utilization in visual language interactions. GCN is a representative model for training representations of attribute graphs. Graph inference trains fixed representations of entities in multiple relational graphs, which are generalized to infer invisible entity relationships during inference. Various approaches are proposed to improve graph reasoning [193], [194].

2) RL-BASED APPROACHES

Multi-agent RL, Imagination-augmented RL, and Language-grounded RL are utilized in Metaverse because RL is suitable for action in a situation without prior learning. Multi-agent RL provides realistic NPCs by causing collaboration and disputes among various agents. Imagination-augmented RL has the feature of rapidly stabilizing without enormous training data, and language-based RL is used for conversation.

Technically, RL is a method to achieve an objective goal by determining the behavior that will receive the maximum reward based on the state received from the environment. It is divided into model-based RL and model-free RL according to the existence of a model for a task. It is also divided into a value-based method and a policy-based method according to the training method. The on-policy method trains an algorithm using the deterministic output of the target policy, whereas the off-policy method indirectly creates and trains a stored distribution. Compensation methods (e.g., episodic memory, world model, and language-based RL) have been proposed to solve the problem of inefficiency and sparse rewards of RL sampling. Furthermore, more efficient approaches (e.g., offline RL and control RL) are emerging to solve fundamental problems (e.g., sample inefficiency, unstable training). Unlike traditional off-policy RL and model-based RL, offline RL uses only pre-collected training data, not online results. Offline RL shows reliable learning with batch training and good performance in a closed-loop environment.

RL methods are steadily growing through knowledge sharing, memory, abstraction, and language bases. The Diversity all you need (DIAYN) model learns useful skills without a reward function, just as humans navigate the environment without supervision [195]. DIAYN acquires skills by maximizing information-theoretic goals using a maximum entropy policy. Laskin *et al.* [196] proposed Contrastive Unsupervised Representations for Reinforcement Learning (CURL) that utilizes the advanced capabilities of raw pixels using contrast learning and out-of-policy controls. Xavier *et al.* [197] proposed a watch-and-help (WAH) model that uses a single demonstration of an agent performing the same task to understand the task's goal and work with a human-like agent to solve a problem.

3) LIFE-LONG LEARNING

Life-long learning is meaningful because it builds experience points over a long period in a sustainable Metaverse. For such life-long learning, a method that effectively memory existing data and use it at an appropriate time is required. Most solutions and services have a constant cycle. In order to apply lifelong learning to Metaverse, it is necessary to consider how to maintain long-term service. On the other hand, the key to life-long learning is how to handle the catastrophic forgetting that most neural net models have.

4) MULTI-AGENT OPTIMIZATION

Relationships between multiple agents are divided into collaborative, competitive, and oracle relationships. In order to effectively utilize these relationships in multi-agents, it is necessary to introduce a mental model (e.g., the Theory of Mind (ToM), intrinsic motivation, and heterogeneous competition). Based on the concepts and experimental results of psychology and neuroscience, there have been attempts to solve the problem of neural networks. In particular, the theory of mind, inductive bias, and intrinsic motivation

were effective methods in embodied visual language interaction [198]–[200].

Will *et al.* [201] used dopamine's reward prediction error theory to explain rich empirical phenomena. They provide an integrated framework for understanding the representation of rewards and values in the brain. They describe that the brain represents possible rewards as a probability distribution rather than a single scalar, and various future outcomes are spontaneously expressed in parallel. Episodic memory tracks functional and structural interactions between brain regions, particularly the hippocampus [202]. Episodic memory, unlike semantic memory, is a descriptive memory that contains information related to the time and place of acquisition. Gradient episodic memory reduces forgetting by transferring previous knowledge to evaluate model training on continuous data [203]. Oudeyer *et al.* [204] explained how psychology and neuroscience conceptualize curiosity and intrinsic motivation as intrinsic rewards for the brain's novelty, complexity, and information scale. Rabinowitz *et al.* [205] designed a mind neuron theory network, ToMnet that uses meta-learning to observe behavior to build agent models. Melhart *et al.* [206] investigated how the emotional mind theory of gameplay influences behavioral recognition, performance, and frustration behavior in facial emotion recognition tasks.

Cooperative multi-agent RL requires a distributed policy but has limitations in coordinating agent behavior in complex environments. Agents reconstruct each other's observations to generate common knowledge in distributed, collaborative multi-agent operations. It is also essential to study how to believe from the perspective of other agents and humans through collaboration [207]. Humans share potential minds (e.g., beliefs) and these social methods are important for recursive reasoning about the potential consequences of other avatar actions.

In order to effectively operate multiple agents, various optimization methods are proposed. Gated propagation networks improve training with attention and gating on graphs that propagate messages between prototypes of different classes and update them in memory of different classes [208]. Multimodal MAML modulates meta-trained prior parameters to enable fast adaptation and improved training on multimodal distributions [209].

5) INTEGRATION OPTIMIZATION

An integrated platform is needed to handle various modals and various events and interactions. Racanière *et al.* [210] proposed an I2A (Imagination-Augmented Agents) for deep RL combining model-free RL and model-based RL. ZEPETO is a platform that is completely provided in the form of a service, and Unity provides more freedom in which developers create the world they want.

6) OPERATION CONSIDERATION

Continuous service through human-centered design and multi-modal interaction is important from an artistic point of view and a scientific point of view based on design philosophy. Meta RL based on few-shot learning is used because real-time performance is poor to analyze service. Graph RL using the structural characteristics of knowledge is also attracting attention. Because planning is essential to perform more complex scenarios, there are many studies on Planning RL. In order to provide stable service on the integrated platform, it is necessary to cope with network bandwidth and failure response physically. In addition, measures against social and politically sensitive issues (e.g., sanctions and hacking) are required.

C. METAVERSE APPLICATIONS

Most of the research on Metaverse is aimed at marketing and investment purposes, emphasizing social utility. The domains where Metaverse is popularly serviced are games and some office applications. Huggett [57] argued that there is a separation between the present reality and virtual reality of virtual heritage and conducted a study of existence and realism within virtual reality. Skarbez [211] introduced mixed reality, real-world modeling, and real-world modeling. For better Metaverse applications, an approach is needed to model and distinguish the differences and the same points between virtual reality and reality.

1) SIMULATION

Metaverse is being serviced in various forms of application. The simulation starts with a game and is also used for social phenomenon research and marketing simulation. Because it has an educational effect through simulation, it is also used for education and museum visits. Simulations depicting real-world tasks are a universally available application in the Metaverse. General simulation is solution-dependent, but the simulation of Metaverse is performed in Metaverse, so it is different from general simulation. Maharg and Owen [212] conducted a study on application simulation for educational purposes. Siyaev and Jo [59] conducted a study on virtual assets and workflow control using aircraft engineer voice commands. In the case of a virtual environment based on the real environment, exaggeration and the intention of the creator can be included in the process of describing the environment in Metaverse. Shi *et al.* [213] studied the difference between the virtual and real environments by evaluating the agreement between the field survey and VR on the landscape.

Gordon *et al.* [214] proposed a hierarchical interactive memory network (HIMN) consisting of a factored set of controllers and operating at multiple levels of temporal abstraction. They also introduce IQUAD V1, which simulates realistic environments of indoor scenes that can be configured with interactive objects. Qiu *et al.* [215] proposed an object-driven visual search algorithm, MJOLNIR (memory-utilized co-hierarchical object learning for indoor room navigation), that learns how to associate objects with prior knowledge. Li *et al.* [216] proposed a MIND (Mental Imagery eNhanceD) module to model the dynamics of the environment and

create objects for a better understanding of the implemented agent. Tamari *et al.* [217] described that natural language in cognitive linguistics (ECL) is inherently executable and driven by metaphorical mappings and mental simulations to schemas learned through hierarchical organization and interaction.

2) GAME

Games are the most common platform in the popularization of the Metaverse. In addition to simply focusing on interest, there are ways to approach to simplify difficult tasks through games. As much as payment and personal information are widely used in Metaverse, a game based on blockchain technology has been proposed [10]. Hide and Seek is a simple yet effective simulation environment for multi-agent work that uses visual representations of objects and scenes from an egocentric perspective [218]–[221]. Baker *et al.* [218] found that agents create a self-supervised automation curriculum that drives new strategies of multiple stages in a multi-agent competitive environment (i.e., hide and seek). Stanica *et al.* [221] Introduced Neurorehabilitation Exercises Using Virtual Reality (INREX-VR), an immersive neurorehabilitation system using virtual reality. They capture real-time user movements in gamified environments and execute complex movements to encourage self-improvement and competition.

3) OFFICE

In order to supplement the sense of space lacking in online solutions in B2B solutions and conferences, some companies introduced and supplemented the offline concept. In this way, the sound occurring in the office and physical elements (e.g., desks and conference rooms) is given a sense of space. Representative examples of office applications include solutions (e.g., Branch, Gather, and Teamflow) and use spatial audio technology to provide speech and footstep sounds according to distance. The Branch is given a game element that offers virtual currency and experience. Teamflow has the advantage of using work-related tools (e.g., file sharing in conjunction).

4) SOCIAL

Because avatars change skin color and gender as desired, they have the advantage of reducing preconceived notions about social discrimination in conversations. These embodied avatars are more advantageous for simulating social problems than in the form of surveys and role-play. Papagiannidis *et al.* [222] conducted a study on the impact of corporate social responsibility focusing on ethical and policy-related issues. De Decker *et al.* [58] introduce the study on the process for solving complex social problems was conducted using Metaverse. Smart *et al.* [223] explained the important characteristics of social change in the Metaverse and future opportunities.

The online requirements for cultural life (e.g., museums and performances) are gradually increasing. Although the limited capacity and time constraints of an offline concert hall are solved, there is still a lack of differences in texture and fine detail that can be felt offline. Tang [224] evaluated the immersive service using Metaverse for educational library orientation. Choi and Kim [45] studied how visitors experience museums by combining beacons and HMDs. Hazan [24] explored how museum social and cultural experiences are evolving.

5) MARKETING

Economic activity is an important content in the Metaverse. It creates an ecosystem that continues economic activity by consuming clothes and goods provided by the production company and producing and selling them with other users. Metaverse is a virtual world to predict the future by reflecting the characteristics of reality realistically. Kaplan *et al.* [5] dealt with how companies see their differences from other social media and utilize their potential. Cagnina *et al.* [6] conducted a study on the business model of a company in Virtual Worlds and Second Life. Papagiannidis *et al.* [25] described Second Life's take on this retail theater experience. ANoghabaei *et al.* [225] covered industry trends in AR and VR technology adoption.

6) EDUCATION

Audiovisual-based education is an important application of Metaverse with a high potential for popularization. Experiential education is important because what you see in writing and how you feel while experiencing it are different. For example, radiation is difficult to experience, so you may preconceive that it is simply dangerous. Through the Metaverse, it is possible to see the educational effects that are considered while analyzing and experiencing radioactivity technically and scientifically in Metaverse [226]. Sung *et al.* [227] compared the level of immersion and three learning outcomes (learning attitude, enjoyment, and performance) based on facial electromyography by comparing marketing students with existing static video presentations and showed that the meta world method is effective in education.

Kemp and Livingstone [4] analyzed the advantages and disadvantages of a multi-user virtual environment for education, and Collins [17] studied how to access, interact, and generate information in higher education. Templeton *et al.* [228] addressed practical and educational considerations for learning teachers, Suzuki *et al.* [9] conducted a study on mutual collaboration in learning IoT. Metaverse is used in PBL, a problem-based learning method as an educational framework [229], [230]. Barry *et al.* [51] evaluated the quality of instruction in the PBL task based on the increase in the number of blinks that made students' emotions unstable and difficult questions. Khan *et al.* [231] proposed safety training for children in the outdoor environment with VR, Kinect sensor, and the Unity game engine. Muhammad *et al.* [232] introduced the effectiveness of handheld marker-based AR in the aspects o performance, motivation, attitude, and behavior for primary school students.

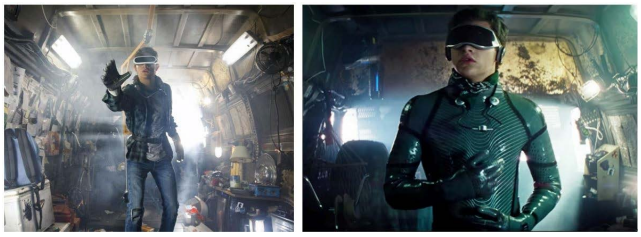

FIGURE 6. Use case of Metaverse movie, Ready Player One.

D. DISCUSSION AND OPEN CHALLENGE

Component models for modal conversion are developed into various forms, from text-to-image conversion to image-to-image translation and video-to-video synthesis. Although the technology for generating the elements of the textual scenario has matured, the integrated research related to the creation of multimodal applications is insufficient. Along with the study of these transformations, there is also a need for studies on E2E learning that simplifies the integration of modules to reduce the complexity of creating multimodal applications. In addition, values, beliefs, attitudes, memories, and decisions are valuable concepts to expand in-depth applications through psychological and neuro-linguistic programming.

V. METAVERSE CASE STUDIES

The utilization of Metaverse in science fiction (SF) is important not only for CGI (computer-generated imagery) but also for UI design. The futuristic UI shown in Minor Report wearable, G-speak, Iron Man HUD, Oblivion, and Enders games provide visual insight for the Metaverse UI. We discuss what technologies were utilized based on the taxonomy proposed in Section 1 in the Ready Player One movie, which is always referred to when talking about the Metaverse. In addition, we do a case study about Roblox, a representative game of the Metaverse. Finally, looking at the recent research results of Facebook Research, the technical possibilities and approaches of Metaverse are summarized.

A. METAVERSE MOVIE: READY PLAYER ONE

1) PHYSICAL DEVICES AND SENSORS

For Head-Mounted Displays, holographic HMDs and goggles-type HMDs were used in the movie as shown in Fig. 6. The holographic HMD is mounted on the neck and plays the role of displaying it on the front. Gloves with sensors that wrap around the hand are used for hand-based input methods. Non-hand-based input methods show in full-body suits that can differentiate and deliver the impact of push, punch, and gunshot. In addition, the shock sensor attached to the chest plays a role in delivering the shock from the virtual world to the human body. As an indirect assistive device, a translucent display tablet used by children at school and a tablet with an expandable screen have emerged. The treadmill comes out as a motion input method, and walking and running are distinguished, and a safe environment is considered by fixing it with a belt.

2) RECOGNITION AND RENDERING

For scene and object recognition, object recognition was used in that the name of the game character and the performance and status of the motorcycle was displayed on the scope. Sound and speech recognition and synthesis were not specifically addressed. Basically, it seems that the recognition and expression of dialogue is a domain with rapid technological development, and it is assumed that it will be in a free state in the near future. Likewise, since it is an animation, they did not deal with specific motion rendering. On the other hand, scene and object generation is used in many places. A holographic map that converts to the actual background while zooming in, the rotation of surrounding buildings before the racing starts, and the effects that look mysterious as a light effect in a dramatic situation are representative examples of scene generation. There were detailed object generation methods (e.g., hair that changes according to feeling), UI interfaces for avatars (e.g., Jarvis), musical instruments played without a player, and textured surface reflections.

3) SCENARIO AND STORY

In a museum, personal photographs, home video recording, surveillance, and nanny cams are expressed in a single multimodal content representation. Metaverse's persona data generation is approaching tall, beautiful, scary, different sex, different species, live-action, and cartoon without restrictions. The protagonist has a virtual best friend who has never met in the real world. His unrealistic colors and shapes (e.g., gray skin, machine body, light source clothing, and fish) of the avatar expresses the strengths of the virtual world well. Another characteristic of the movie is the thorough anonymization between the avatar and real self. What happens in the Metaverse is considered as a form that excludes direct influence on the real world. Representative virtual NPCs are shown as a simulation curator who helps introduction and events and an avatar of a creator who progresses the story. Multimodal entity linking is used to create 3D virtual experiences with personal photographs, home video recording, surveillance, and nanny cams. In the case of scenario generation, an important theme runs through the entire film. There are many sub-quests (e.g., collecting coins in the event space, planet12) and the main quest that continues the story.

The entire scenario line features the death of a respected game creator, as well as massive rewards for Easter eggs as missions. The creator's avatar guards three magic gate keys in various places, and the avatars do an adventure to find the keys for the Easter eggs, which rule in Metaverse. The avatars carry it out in the Metaverse through missions that require a reasonable level of common sense and reasoning. It makes you more immersed in the scenario by showing things that are difficult to experience in reality (e.g., announcing the start from the flames of the Statue of Liberty and flying subways). The scenario is populated with the visual experience similar

to the real ones in the virtual environment while converting the viewpoint and speed. Scenario evaluation shows whether there is any contradiction while combining the events before and after about the relationship between the creator and his girlfriend.

4) USER INTERACTION

Although language and multimodal interaction are not specifically mentioned, it assumes an organic combination with no synchronization and awkwardness in processing 3D stereoscopic screens and conversations as multi-task interactions. Showing the number of kills and damage by overlaying them on the gun, passing money, and throwing objects in the air shows the advantages of embodied interaction. A car is carried as small as a car key and be put into virtual space as an inventory concept. It gradually expands from a small icon to a real-sized car when taking out the car, giving visual pleasure. User interaction with NPC is also considered. The AI robot helps the protagonist by using a search system in the museum. Immersive OASIS connection video plays a role in distinguishing the real and the virtual. The last motion to exit the Metaverse is used as a pose to take off the goggles.

5) METAVERSE APPLICATIONS

Basic simulations (e.g., ballet, boxing, piano, dance, and tennis) appear at the beginning. The opening video shows that things that are difficult to experience in ordinary life (e.g., hang gliding, waves in Hawaii, skiing at the pyramids, climbing with Batman and Everest, a planet-sized casino, divorce, and marriage) are possible. In the Metaverse, Gregarious games, Minecraft, and 3D pinball are mini-games, and avatars receive reward coins according to their level and risk. Avatar acquires coins when a car and a person breaks in Metaverse, but visual effects and damage realistically apply to vehicles collide. Through a scoreboard with rankings in the Metaverse, the intermediate process and results of the game are shared.

Regarding the Office, we expect an organizational approach of a company when looking at the employees of IOI companies who work in an independent space in a company with the shape of an avatar. It also talks about the possibility of an organized group with a commercial goal as an interesting company appearing inside the Metaverse and going to a racing game as if it were a job. From a social point of view, class according to grade, disconnection from children due to game participation, and side effects of anger caused by immersive game participation are described.

There is a view of blocking from the real world, which is shown by anonymizing names with numbers instead of names. On the other hand, some example shows that communicate through the interface between the real world. Even in the Metaverse world, a marketing singularity sells offline items (e.g., suits) and is appropriately used in game items. Although it is in the Metaverse that mimics the real world, unrealistic control function items in the Metaverse (e.g., the time turning item) are also used in a balanced way that does not threaten the world view. The appearance of education suggests that the form of education will not be much different from what it is now, although a translucent display window tablet is used.

FIGURE 7. Use case for Metaverse game, Roblox [1].

6) DISCUSSION AND OPEN CHALLENGES

Player Ready One shows negative aspects of the Metaverse (e.g., surrogate exam, taste cheating, and mirroring). The problem of over-addiction is explained in the appearance of upgrading a suit with the money to be paid for the rent due to the virtual world and excessive immersion. Metaverse is based on separation from reality but depicts the fact that virtual damage is done to the real world. Finding the owner of the avatar in the real world and jumping out the window in anger over defeat are mental problems in the immersive Metaverse. The appearance of falling off a chair and falling backward is expressed as an example in which the Metaverse inflicts real physical damage.

Metaverse implementation is not described in detail in the movie except for the concept of life-long learning for the museum scene because it is a technical detail. By visualizing the avatar as a hologram in the real world, they showed it is possible not only in the forward direction (i.e., from reality to the Metaverse) but also in the reverse direction (i.e., from the Metaverse to reality). Eating, sleeping, and bathroom breaks are seen as new expandable possibilities in the sense that they are not done in the Metaverse.

B. METAVERSE GAME: ROBLOX

Roblox served by two-thirds of 9-12 years old in the United States and is a representative game of Metaverse with an MAU of 150 million [1]. Roblox is also used to develop simulations of urban environments to describe experiences that incorporate the realization of virtual paths to the city's sculptural heritage in the classroom, as shown in Fig. 7. Students were able to understand and integrate Santa Cruz's sculptural heritage to create their own interactive world with Roblox [2]. Creatively interpreting legacy in both formative and programming is a good approach. Although norms between education and entertainment have often been regarded as two separate worlds, Roblox is used as an educational tool in the classroom from the perspectives of motivation, problem-solving, and STEM [3].

1) PHYSICAL DEVICES AND SENSORS

Roblox supports Oculus Quest 2 and HTC Vive for 3D HMD. VR has a gyroscope, display screen, and built-in audio

system that provides a VR experience as an independent device that is different from the Google Cardboard method using a smartphone screen. The Quest 2 is compatible with PlayStation VR and does not require a gaming PC, but a VR-capable PC is required to run Roblox. It is the cheapest VR head for Roblox, and it has a vast resolution of 1832 × 1920 pixels per eye, but the front of the device is heavier, which is inconvenient. With full VR support, users play games like Skyrim, Half-Life: Alyx, and No Man's Sky in Roblox. The HTC Vive Pro has a resolution of 1440 × 1600 pixels and has a built-in audio system. There is additional padding throughout the device to share weight and balance, but the HTC Vive Pro only works via DisplayPort. Roblox supports VR devices and supports VR for some games, but it is still limited. In addition, most devices are limited to HMDs, so they lack the versatility of a tactile and pressure sensor that a normal Metaverse would consider supporting. Since the main customers are a young generation, they seem to focus on simple forms over complicated ones.

2) RECOGNITION AND RENDERING

The SW used in Roblox is Lua and Roblox studio. Lua is a small interpreter language with a capacity of several hundred tens of KB. The script is a programming language that can be executed line by line and is a tool that creates events that occur in the game, physics engine, text output, and screen effects. It was developed with the goal of being a lightweight scripting language with a clean syntax that is easy to embed into C/C++ programs.

Roblox Studio is the official free utility software for creating custom games for Roblox. Users configure various game worlds and servers (e.g., mini-games, obstacle courses, and role-playing stories). Its characteristic is that even low-quality games are made by own hand and enjoyed with friends. Because operation on low-end devices (e.g., mobiles) cannot be played in high-end games, they have an experience with lag, skin color errors, dialogue control errors, airplanes malfunction, etc. UnityML has the advantage of being able to link and use various recognition methods in a 3D environment, but Roblox is often composed of simple and lightweight forms, so that part is insufficient.

3) SCENARIO AND STORY

There are many young users, so the game effects and scenario complexity are low, but it is diverse and novel. It is an online game creator system in which most of the content is produced by amateur game creators. User-generated content is an avatar accessory created by a user, and it is an item that can be created when a user has a reputation within the community and is an expert who handles modeling programs well. When users satisfy the conditions and get Roblox's certification, users get permission to create items using mesh, and users have to pass Roblox's review. It doubles the elements of the game with limited items but also creates a counterfeit UGC and copies unique items to bring about a deflationary effect. It is also used for the misuse of items that can disguise the character in the Roblox game.

On the other hand, the strength of Roblox is that various users can easily create new games. Although it is relatively simple, it presents a new perspective with various and novel approaches. However, each game is centered on a single story and lacks the depth of the story because it does not have an elaborate plot. Sometimes, the story is similar, and the story development is not stable. When there is an authoring tool (e.g., multimodal story generator) to make and evaluate plots, users easily create more in-depth content and games.

4) USER INTERACTION

The Metaverse trading system supports user exchange with other players for dollars, so a connection with the real world is also considered. Premium service provides a differentiated service that makes shirts and pants, sells them free, and sells them at a price. It also reinforces the interaction by providing an online hangout-concept space called a separate party place. There are various auxiliary methods (e.g., facial expressions, clothes, motions, and words) to express their feeling in Roblox. However, there is room for improvement in real-time and tactile interactions. Special attention is required for interaction because children spend a lot of time. It supports multiple languages but has a low level of translation quality.

5) METAVERSE IMPLEMENTATIONS

One of the most problematic for Metaverse commercialization is stable operation, especially 3D rendering, for many concurrent users. From an operational point of view, there are problems with hacking, extortion, and server down. Management and efforts are in place to ensure user safety (e.g., prevent profanity, review on image uploads, parents prohibit chatting, more than 1,600 administrators), but as the scale grows, the number of users' improper behaviors increases. Games administrator build their own reporting systems for these shortcomings and sanction them. On the other hand, excessive restrictions and privacy authority are also a problem. There is a privacy issue where the management can censor personal messages and know the current location.

6) METAVERSE APPLICATIONS

Roblox is a game playground. Since there are not many games that elementary school students can play easily and comfortably, it can be seen as an imaginative game that can be seen in playgrounds. Since game items and passes are possible to break the game balance, the balance is important for commercialization. A concert called One world together at home is also opened as an application. It is used as a tangible connection medium to generate revenue through the production of ZEPETO items and to deliver from the real world to the virtual world using Roblox currency. There are phenomena that are seen in general society, hyperinflation following the abolition of Ticks.

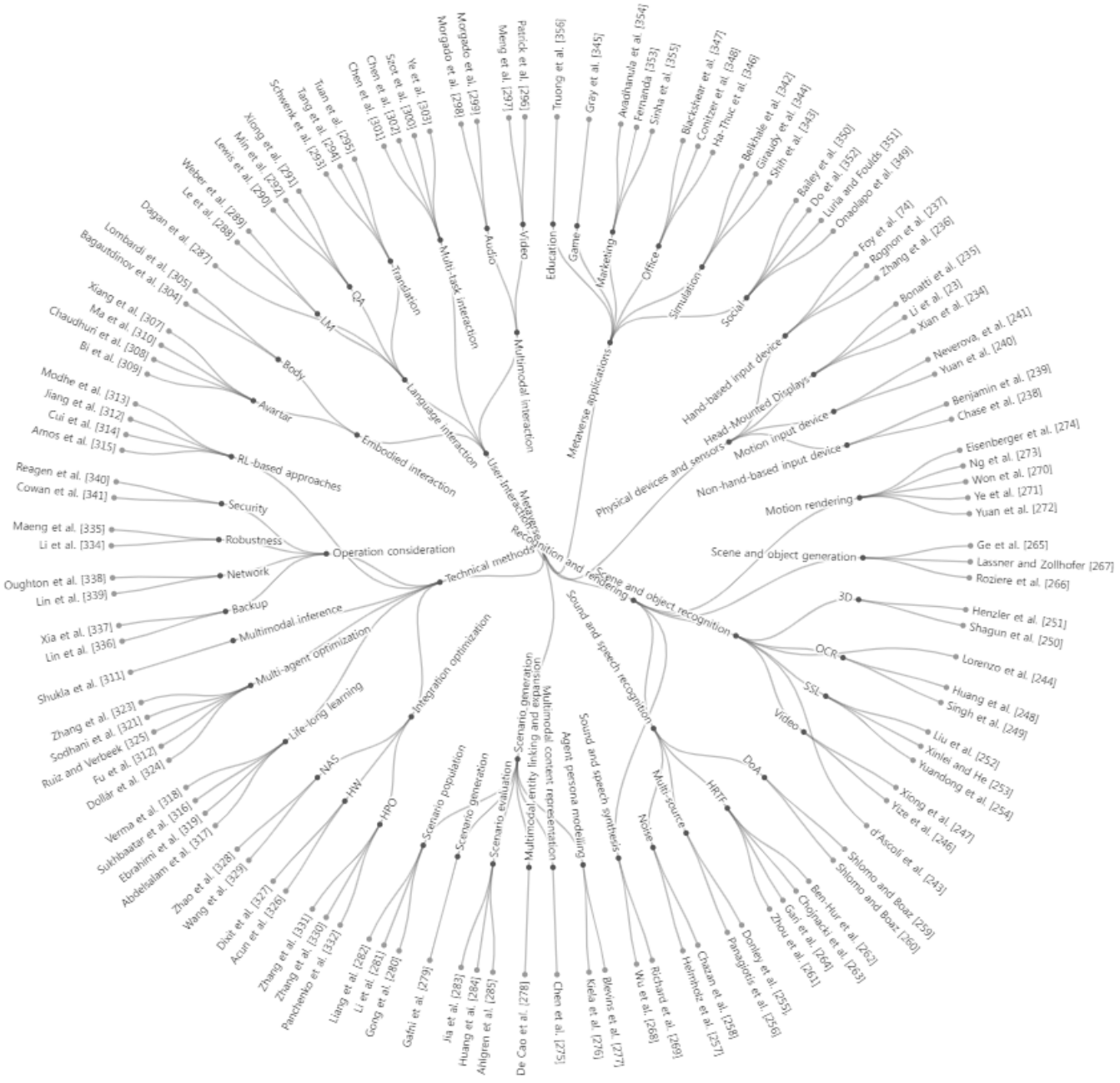

FIGURE 8. Recent researches of Facebook research.

7) DISCUSSION AND OPEN CHALLENGES

Roblox supports VR, but non-VR games account for a significant portion, and there is a possibility that it will develop into a more advanced form based on a large number of subscribers. Roblox is well known to the younger generation, so children can learn to code and make friends by taking Roblox coding classes and camps. However, there is a problem that it is difficult to check all the contents because there is 50 million game content despite the overall user acceptance level. This management problem is problematic in that the primary user class is relatively lower age.

C. METAVERSE RESEARCH: FACEBOOK RESEARCH

Based on the papers published in Facebook Research from January to June 2021, we classified each paper into the taxonomy defined in Section 4 and summarized our approach in terms of Metaverse utilization, as shown in Fig. 8.

1) PHYSICAL DEVICES AND SENSORS

a: HEAD-MOUNTED DISPLAYS

One of the hallmarks of Metaverse using a head-mounted display is that it sees the world from an egocentric perspective. Most video processing uses third-person video

data sets, so egocentric video data is not enough. Third-person view data is not directly available in the Metaverse due to the inconsistency of the viewpoints, so an approach that transforms it into an egocentric video model is required. Li *et al.* [233] generated a model that exploited knowledge distillation loss during pre-training to obtain both the scale and diversity of third-person video data, as well as representations with prominent egocentric properties.

Xian *et al.* [234] presented a method for learning a spatiotemporal neural irradiance field for a dynamic scene that enables preview rendering of the input video. Using the scene depth estimated in the video, they constrained the time-varying geometry of the dynamic scene representation and presented a single global representation of the contents of individual frames. Generating expressive camera motion for autonomous flight technology is difficult because it requires editing of several control parameters that are not intuitive for users. Bonatti *et al.* [235] developed a data-driven framework for editing complex camera position parameters in semantic space. They constructed a semantic control space by analyzing the correlation between technicians based on the study of filming guidelines and human perception.

b: HAND-BASED INPUT METHODS

On a physical keyboard, the resistance of the keys prevents erroneous input, but in Metaverse, it is needed to isolate spurious input events when typing with a virtual keyboard. Foy *et al.* [74] showed three alternative co-activation detection strategies with high accuracy. They developed StickyPie, a marking menu technology that enables scale-independent marking input by estimating intermittent landing positions. They identified issues inherent in eye movement control and current eye-tracking hardware, including erroneous selection activation, while reducing workload and eyestrain.

Natural hand manipulation is a task that requires complex finger manipulation to adapt to the shape and task of an object. Zhang *et al.* [236] proposed a generalizable hand-object space representation combining voxel occupancy and global object shape with local geometric details to the nearest sample. Hand social contact is essential for social interaction and communication and reduces anxiety and loneliness. Rognon *et al.* [237] Introduced mediated social contact that conveys indescribable emotions (e.g., love, empathy, reassurance), allowing devices to transmit haptic signals and physically interact at any distance.

c: NON-HAND-BASED INPUT METHODS

Research on input devices using wrist motion without directly attaching to the hand is also increasing. With the growing interest in vibrotactile feedback in wearable wristband devices, Chase *et al.* [238] used information transfer as a metric to explore the signal variation space within a single vibrotactile actuator (e.g., frequency, amplitude, and modulation). Typical control systems rely on digital on/off control to limit the degrees of freedom available when designing haptic experiences, allowing only inflate/decrease at a set rate. Stephens-Fripp *et al.* [239] presented an alternative system in which analog control of the pneumatic wave profile can be used to determine the optimal wave profile. The attack and release profile have been altered to create a more pleasant pulsating sensation at the wrist and a more lasting sensation of transmitting movement around the wrist.

d: MOTION INPUT METHODS

To accurately estimate 3D human movement, both kinematics (i.e., body movement without physical force) and dynamics (i.e., movement with physical force) must be modeled. Yuan *et al.* [240] presented a SimPoE, a simulation-based approach for 3D human pose estimation that integrates image-based kinematic inference with physics-based dynamic modeling. To obtain accurate pose estimates, a meta-control mechanism was used that dynamically adjusts the character's dynamic parameters according to the character's state. Neverova *et al.* [241] jointly learned the geometry of several categories of deformable objects to learn integrated dense pose predictors for several categories of related objects. It has symmetric inter-category periodic consistency and a new asymmetric image-category periodic consistency and has improved performance over methods for 3D shape matching without manual annotation of inter-category correspondences.

2) RECOGNITION AND RENDERING

Lucas and Kozary [242] focused on the basics of teaching computers to think like humans when making decisions about visual content that are most interesting and important to human viewers. Computers see colors as numbers rather than meaningful parts of an image, and textures see numbers rather than meaningful hard and soft parts of an image. Some parts are similar to human perception, but there are also other parts, so the difference between humans and computers is an important research field.

a: SCENE AND OBJECT RECOGNITION

The hard inductive bias of CNNs allows for sample-efficient learning, but at the expense of potentially lower performance limits. Vision Transformers (ViTs) rely on more flexible self-attention layers and perform better than CNNs in image classification. However, expensive pre-training on large external data sets or distillation of pre-trained convolutional networks is required. d'Ascoli *et al.* [243] introduced gated positional self-attention (GPSA), a form of positional self-awareness equipped with soft convolutional induced bias. The use of cropping can bias large objects to be clipped or omitted, as described in Lorenzo *et al.* [244] proposed a new crop recognition bounding box regression loss (CABB loss) that facilitates prediction to match the visible part of the cropped object. In response to the disproportionate distribution of object sizes, they introduce a new data sampling and augmentation strategy that improves generalization across scales. Cheng *et al.* [245] updated

the standard evaluation protocol, for instance, and panoptic segmentation tasks by proposing Boundary AP (Average Precision) and Boundary PQ (Panoptic Quality) metrics, respectively, based on Boundary IoU for image segmentation evaluation.

The spatially deformed spatial resolution of the retina is utilized for foveated video compression for immersive video requiring large bandwidths of high spatial and temporal resolution. Yize *et al.* [246] proposed FED (Foveated Entropic Differencing), a Full Reference (FR) centric image quality evaluation algorithm for centric video compression. Xiong *et al.* [247] presented a multi-view pseudo-labeling approach using complementary views in the form of shape and motion information for semi-supervised learning in the video. By acquiring pseudo-labels from unlabeled videos, more robust video representations were learned than purely supervised data.

It is also necessary to use text information as well as video. Huang *et al.* [248] proposed Multiplexed Multilingual Mask TextSpotter, an E2E approach, for end-to-end education and scalable multilingual multi-purpose OCR system. They kept the integration loss of performing script identification at the word level and processing different scripts with different recognition heads while simultaneously optimizing script identification and multiple recognition heads. In the past, scene text-based inference separated from OCR systems was difficult due to the lack of ground-truth text annotations or scene text detection and recognition datasets for real images. Singh *et al.* [249] introduced a TextOCR for detecting and recognizing scene texts of arbitrary shape with 900k annotation words collected from real images in the TextVQA data set.

Since the Metaverse assumes a 3D environment, many 3D-related skills (e.g., fast rendering and few-shot learning) are required. Sodhani *et al.* [250] proposed an open-source OpenNEED consisting of a large-scale, high-frame-rate non-eye (head, hand, and scene) and eye (3D gaze vector) data set. They proposed a robust eye tracker design considering non-eye sensors to study the relationship of head, hand, scene, and gaze and apply spatiotemporal statistics to gaze estimation. Henzler *et al.* [251] proposed a new neural network called warp-conditioned ray embedding (WCR) that focuses on training a model on multiple views on a large collection of object instances to learn a deep network that reconstructs in 3D given a small number of images. Ren *et al.* [75] introduced WyPR, a weakly supervised framework for point cloud recognition that requires only scene-level class tags as a director. They proposed to solve jointly by combining point-level semantic segmentation, 3D proposal generation and 3D object detection, and self- and cross-task coherence loss prediction.

Liu *et al.* [252] proposed an Unbiased Teacher to identify the pseudo-label bias problem of SS-OD (Semi-Supervised Object Detection). It is a simple but effective approach to train students and progressively develop teachers mutually beneficially jointly. Chen and He [253] used negative sample pairs, large batches, and momentum encoders to avoid solution decay and showed that the gradient stopping operation plays an essential role in preventing decay. Tian *et al.* [254] studied the nonlinear learning dynamics of uncollated SSL in a simple linear network where SSL with only positive pairs avoids expression decay. They investigate conceptual insights into how the disjoint SSL method learns, how to avoid expression collapse, and how several factors (e.g., predictor networks, stationary gradients, exponential moving averages, and weight reduction) work.

b: SOUND AND SPEECH RECOGNITION

Metaverse runs in a variety of places, from a relatively quiet house to a space where a variety of people gathers. Donley *et al.* [255] proposed Linearly Constrained Minimum Variance (LCMV), an automated solution for multi-channel signal enhancement to improve voice communication in a noisy environment. They use the beamformer to estimate the relative source contribution of each source in the mixture and then used to weight statistical estimates of the spatial properties of each source used for the final separation. It allows instant selection of desired and undesired sources. Furthermore, it improves multi-channel speech enhancement for dialogue, aiming to extract clear speech from a noisy mixture using signals captured by multiple microphones. Panagiotis *et al.* [256] applied a graph neural network (GNN) to find the spatial correlation between various channels and integrated it into the embedding space of the U-Net architecture with the graph convolution network (GCN). Helmholz *et al.* [257] introduced Real-Time Spherical Array Renderer (ReTiSAR) to analyze the sensor's own noise propagation through the processing pipeline. The instrumental evaluation confirmed the strong global impact of various arrays and rendering parameters on spectral balance and the overall level of rendered noise. They determined the audible threshold of coloring artifacts during head rotation for various array configurations in a perceptual user study. Helmholz *et al.* [257] applied binaural rendering of a spherical microphone array signal to increase the SNR of the rendered signal by up to 9 dB with some array configurations with larger radii and spherical harmonic order four or higher microphones. Chazan *et al.* [258] presented an integrated network for speech separation of an unknown number of speakers and presented a noise and reverberation dataset for five speakers.

Research on high-quality surround sound audio is based on a fixed position of the recording microphone in general, such as in a movie theater. However, in Metaverse, users can change their listening position as they run, spin, or various body changes. Birnie *et al.* [70] proposed a method for binaural playback of microphone recordings in a virtual application in which one's body freely moves beyond the recording location. They integrate near, and far sources in an extended virtual environment and better reproduce the intensity and binaural room impulse response spectrum of the near environment.

Early room reflection estimation is an important task in audio signal processing, along with beamforming, source separation, room geometry inference, and spatial audio applications. Shlomo and Boaz [259] proposed a solution for blind estimation of reflection amplitudes using iterative estimators based on maximum likelihood and alternating least squares. Blind estimation of direction of arrival (DOA) and delay of indoor reflections due to reverberation is useful for a wide range of applications, but conventional methods detect only a few reflections. Shlomo and Boaz [260] proposed PHase ALigned CORrelation (PHALCOR) for estimating early reflex delay and DOA blinding.

Head-Related Transfer Function (HRTF) is used to simulate external sound by measuring the sound source's spectrum in three-dimensional space. HRTF individualization enables realistic and immersive spatial audio rendering in Metaverse. Zhou *et al.* [261] identified the lowest spectral distance error by exploring the range of HRTF predictability using a deep neural network with a 3D ear shape as input. In practice, binaural reproduction is also affected by HRTF, along with truncation errors that detrimentally affect the perception of the reproduced signal. Because pretreatment of HRTF by ear alignment prevents effective recognition, Ben-Hur *et al.* [262] presented a method for integrating preprocessed ear-aligned HRTFs into the binaural regeneration process. HRTF is the key to audio spatialization. However, it cannot produce sufficient sound output levels at low frequencies (below 300 Hz) while maintaining an omnidirectional pattern. To address this problem, Chojnacki *et al.* [263] proposed a new design to overcome the limitations of this low-frequency range at higher frequencies. Gari *et al.* [264] analyzed and rendered multi-channel RIR (Room Impulse Response) by parameterizing the sound field as a series of plane waves for the Spatial Decomposition Method (SDM). They reduced the unnatural arrival direction diffusion of late reflections by spatial clustering of reflections in the post-processing and solved the whitening problem of late reverberations with a binaural RIR corrected equalization method, RTMod+AP.

c: SCENE AND OBJECT GENERATION

Ge *et al.* [265] proposed DoodlerGAN, a generative part-based GAN (Generative Adversarial Network) that generates creative and high-quality images to generate invisible configurations of new part shapes. They also introduced two creative sketch datasets: Creative Birds and Creative Creatures. Aiming to increase the resolution and level of detail within super-resolution images, Roziere *et al.* [266] utilized an evolutionary method to improve NESRGAN+ by optimizing noise injection at inference time. They proposed Diagonal CMA to optimize the injected noise according to a new criterion that combines quality assessment and realism. Lassner and Zollhofer [267] proposed Pulsar, an efficient sphere-based differential rendering module that is fast, modular, and easy to use. It avoids topological problems by using spheres for scene representation. It uses an efficient differential projection operation and neural shading to alleviate topology inconsistency problems, high memory footprint, and slow rendering speed.

d: SOUND AND SPEECH SYNTHESIS

The computational complexity of the transformer increases twofold with sequence length, making it impractical for many real-time applications. Wu *et al.* [268] proposed an efficient transformer-based acoustic model with constant speed regardless of input sequence length for streaming speech synthesis applications. They used the Emformer network to predict frame rate spectral characteristics in streaming and WaveRNN neural vocoder to generate the final audio by taking the predicted spectral characteristics. They demonstrated consistent performance, low latency, and low real-time performance over various utterance lengths. Richard *et al.* [269] presented a neural rendering approach for binaural sound synthesis that generates spatially accurate binaural sound in real-time. They proposed end-to-end neural binaural sound synthesis that outperforms DSP-based methods in a perceptual study and a qualitative evaluation.

e: MOTION RENDERING

Control strategies for physically simulated characters performing two-person competitive sports (e.g., boxing and fencing) are used as a reference for effective motion rendering in the Metaverse. Won *et al.* [270] developed a learning framework for generating control policies for physically simulated athletes with many degrees of freedom. They presented a control policy learned from a framework that generates both tactical and natural behavior. Ye *et al.* [271] proposed a learning-based approach that infers an object's 3D shape and poses from a single image and learns from a collection of atypical images supervised only by the segmented output of an off-the-shelf recognition system (i.e., shelf supervision). They inferred the volume representation of standard frames together with camera poses. After that, they performed shape-pose decomposition and instance-by-instance reconstruction of image collections in more detail. Yuan *et al.* [272] proposed a STAR that performs self-supervised tracking and reconstruction of dynamic scenes with rigid motion in multi-view RGB video without manual annotation. By decomposing into two component parts and encoding each into its own unique neural expression simultaneously, the dynamic scene is reconstructed as a single solid object in motion. They also jointly optimized the parameters of the two neural luminosity fields and a set of fixed poses that align the two fields in each frame. Ng *et al.* [273] studied body motions for 3D hand shape synthesis and estimation in the area of conversational gestures based on the assumption that body movements and hand gestures are strongly correlated in non-verbal communication environments. Hand prediction model generates a 3D hand gesture with only the 3D motion of the speaker's arm as input. Eisenberger *et al.* [274] proposed a neural network architecture NeuroMorph that takes two 3D shapes as input and generates them at once in an end-to-end learning

method. It is in a fully unsupervised manner without manual correspondence annotation. By combining graph convolution with global feature pooling to extract local features, geodesic lines are approximated in this shape-space manifold to produce realistic deformations.

3) SCENARIO AND STORY

a: MULTIMODAL CONTENT REPRESENTATION

In the task of retrieving linked query images from a database, Chen *et al.* [275] proposed to express an image as a constituent object based on the intuition that the finest detail of manipulation is often at the object level. They introduced an object-embedding framework for OE-SIR (Spliced Image Retrieval) using object detectors and a teacher-student model to localize object regions.

b: PERSONA DATA GENERATION

Kiela *et al.* [276] introduced Dynabench, an open-source platform that runs in a web browser and supports the creation of human-in-loop model datasets for dynamic model benchmarking. With Dynabench, data set creation, model development, and model evaluation inform each other directly, making it a more powerful and informative benchmark. Current models for Word Sense Disambiguation (WSD) are human-level performance in global WSD metrics but lack data to model and evaluate rare senses. Blevins *et al.* [277] established criteria for FEWS using knowledge-based neural WSD approaches and better captured rare sensations in the WSD dataset with a model further trained with FEWS.

c: MULTIMODAL ENTITY LINKING AND EXPANSION

Context and entity affinity are mainly captured via vector dot products, potentially missing fine-grained interactions between them, requiring large memory footprints to store dense representations. De Cao *et al.* [278] proposed GENRE to generate unique names for each token in a left-to-right autoregression method and search for entities according to context. It directly captures the relationship between context and entity name, effectively cross encoding both and greatly reducing memory footprint because it scales with the lexical size rather than the number of entities.

d: SCENARIO GENERATION

For the motion transfer task between the one dancer and the target person, Gafni *et al.* [279] proposed a model to re-animate a single image with an arbitrary video sequence. They combine three networks: a segmentation mapping network, a realistic frame-rendering network, and a face enhancement network.

e: SCENARIO POPULATION

Data augmentation methods experience distribution shifts and consequently degrade the performance of non-augmented data during inference. Gong *et al.* [280] used a saliency map to detect important regions in the original image and preserved these information regions while augmenting them. Because moments extracted from instance normalization and position normalization roughly capture the style and shape information of the image, Li *et al.* [281] proposed Moment Exchange, an implicit data augmentation method that encourages models to utilize moment information in recognition models as well. Knowledge Distillation (KD) tends to make inconsistent predictions when the data distribution changes slightly, so a method is needed to apply it to low-resource (both memory and computational) platforms. Liang *et al.* [282] proposed MixKD, a data-agnostic distillation framework that utilizes a simple but efficient data augmentation approach to give the resulting model stronger generalization capabilities.

f: SCENARIO EVALUATION

Jia *et al.* [283] systematically studied whether the extent visual information (i.e., objects and contexts) contributes to understanding human motives to analyze how visual information easily recognizes human intentions behind social media images. They introduce Intentonomy, an intent dataset consisting of 14K images covering a wide range of everyday scenes to study the present intentions. When training intent classifiers, they performed additional studies to quantify the effects of attending object and context classes and textual information in the form of hashtags. Huang *et al.* [284] implemented a post-processing step with simple modifications to the standard label propagation technique in the initial graph-based semi-supervised learning method. A cyber-physical digital twin is a simulation of a non-software (physical) system, which has recently received much attention, but its cyber-cyber response is relatively overlooked. Ahlgren *et al.* [285] measured the practical impact on digital twins' design, implementation, and deployment as conceptually true twins by simulating other software systems.

4) USER INTERACTION

Speech recognition-based natural language dialog is the basic medium of user interaction. Recently, BlenderBot 2.0 [286] was proposed based on two studies: Internet search engine-based generation and long-term memory integration, as shown in Fig. 9. The LM-based dialog generation model has the hallucination problem of generating plausible sentences that are factual. To prevent this problem, searching using the Internet and generating a final response based on the searched information was proposed. In addition, in most conversational studies, many short conversations (typically 2-15 turns) consist of a single conversation session because the dialogue engine gives scenario-specific answers rather than responses based on long-term memory. The proposed model provides improved search capabilities with the ability to summarize and recall previous conversations. However, since it is a model with an open dialogue that can expose personal behavior (e.g., long-term memories and the

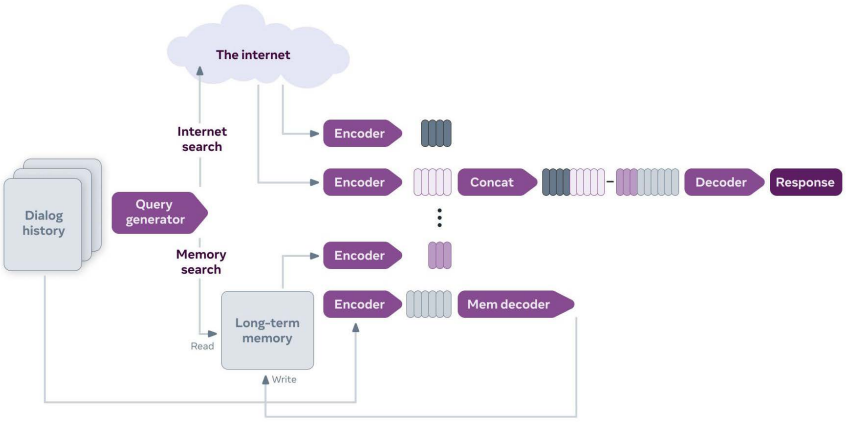

FIGURE 9. Facebook Chabot Blender bot 2.0, which utilize memorization and internet, search results [286].

speaker's personal interest), careful attention is required in management.

a: LANGUAGE INTERACTION

Reference games illustrate the functional use of language for communication and provide a basic learning environment for neural agents. Languages are inherently biased by the underlying capabilities of agents. Dagan *et al.* [287] introduced the Language Transmission Simulator to model agent populations' cultural and architectural evolution. They emphasize the importance of studying basic agent architectures and propose coevolution of languages and agents in the study of language emergence. With the recent development of LM, it is widely used for various tasks of natural language and various modals. Recurrent Neural Network Transducer (RNN-T) is a famous method in automatic speech recognition due to its simplicity, conciseness, and general transcription, but it lacks an external language model and is more vulnerable to rare long-tail words (e.g., entity names). Le *et al.* [288] proposed RNN-T to model intractable rare WordPieces by injecting additional information into the encoder and using alternative letter pronunciations. Deep fusion with personalized language models for stronger biasing. Weber *et al.* [289] considered language modeling as a multi-task problem, combining three studies: multi-task learning, linguistics, and interpretability, to analyze the generalization behavior of language models in Negative Polarity Items (NPIs).

QA is the most basic solution for communicating with NPCs in the Metaworld. Annotated data sets are difficult and expensive to collect and rarely exist in languages other than English. That is the reason it is hard to build a QA system that works well in other languages. Lewis *et al.* [290] proposed a multi-dimensionally ordered extractive QA evaluation benchmark MLQA. Xiong *et al.* [291] proposed a simple and efficient multi-hop dense search approach to answer complex open-domain questions, achieving state-of-the-art performance in two multi-hop data sets, HotpotQA and multi-evidence FEVER. Min *et al.* [292] proposed a model to

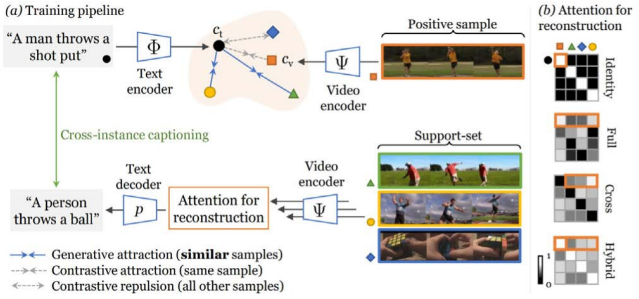

FIGURE 10. Multimodal training with videotext representation learning [296].

build a system that can predict correct answers in open QA that receives natural language questions as input and returns natural language answers while meeting strict disk memory budgets. Memory budgets encourage agents to explore a balance between storing parameters for large and redundant search corpora and large training models.

Multilingual support is required to compose a natural interface while covering a wide range of Metaverse. Because the common language (e.g., English) has limitations for fluent communication, multilingual translation is required to provide a natural interface in other languages. Schwenk *et al.* [293] presented a multilingual sentence embedding-based approach to automatically extract parallel sentences from the content of Wikipedia articles in 96 languages. Other modalities tend to generate similar decoder representations and preserve more information in pre-trained text translation modules. Tang *et al.* [294] proposed a parameter sharing and initialization strategy to enhance information sharing between tasks. It is a new attention-based regularization for encoders and an online knowledge distillation method to improve knowledge transfer. The quality assessment aims to measure the quality of translated content without access to reference translations. Tuan *et al.* [295] proposed a method that does not rely on examples from human commentators but instead uses synthetic training data.

b: MULTIMODAL INTERACTION

Noise contrast learning for videotext representation learning increases the similarity of representations of pairs of known samples and repels all other representations. Patrick *et al.* [296] proposed a method to mitigate this by using generative models to push these related samples naturally, as depicted in Fig. 10. The captions of each sample were reconstructed as weighted combinations of the visual representations of other supporting samples. It is difficult to learn the grounding of each word due to noise and the presence of words that cannot be visually meaningfully grounded. Meng *et al.* [297] presented a jointly trained model architecture for controlled trace generation and controlled caption generation. They proposed a local bipartite matching (LBM) distance measurement that compares two traces of different lengths to evaluate the quality of the generated trace. Because audio and video signals are not always informed of each other, audiovisual correspondences often result in false positives. It optimizes the weighted-contrast learning loss and lowers its contribution to the overall loss. Morgado *et al.* [298] optimized the instance identification loss with a soft target distribution that estimates the relationship between instances. Morgado *et al.* [299] optimized visual similarity rather than simple cross-modal similarity using SS based on contrast learning with cross-modal audio and visual recognition.

c: MULTITASK INTERACTION

Szot *et al.* [300] proposed a simulation platform for training virtual robots in interactive 3D environments and complex physics-based scenarios. Experimental results showed that flat RL policy suffers from HAB (Home Assistant Benchmark) compared to the hierarchical policy, hierarchical structure with independent technology suffers from takeover problem. In audiovisual exploration, agents use both sight and sound to move through complex and unmapped 3D environments intelligently. Chen *et al.* [301] showed how to operate at a fixed granularity of agent behavior and rely on simple iterative aggregation of audio observations, as shown in Fig. 11. It uses waypoints that are dynamically set, and end-to-end learned within the search policy. Acoustic memory provides a structured and spatially based record of what the agent hears as it moves. Recent work on audiovisual navigation assumes a continuously audible target, and the role of audio in announcing the target's location is limited. Chen *et al.* [302] introduced semantic audiovisual exploration in which objects in the environment make sounds consistent with their semantic meaning (e.g., flushing toilets, creaking doors) and in which acoustic events are sporadic or short-duration. They proposed a converter-based model for handling this new semantic AudioGoal task by incorporating an inferred goal descriptor that captures an object's spatial and semantic properties. Persistent multimodal memory allows the target to be reached even after the acoustic event has stopped. ObjectGoal Navigation (OBJECTNAV) is the task of an agent navigating object instances in an invisible environment, which degrades performance due to overfitting and sample inefficiency. Ye *et al.* [303] integrated the learned components and motivated methods that operated on explicit spatial maps of the environment and reactivated the general learning agent by adding auxiliary tasks and navigation rewards.

d: EMBODIED INTERACTION

It's unclear how to optimize the layout of 3D UI controls for body and aerial interactions. Li *et al.* [72] evaluated the performance and limitations of a non-dominant fixed 3D UI in a VR environment through a two-handed pointing study. It has been demonstrated that targets that appear closer to the skin (i.e., located around the wrist placed on the inside of the forearm) can be selected faster than targets that are further away from the skin (i.e., around the elbow on the side of the arm). Bagautdinov *et al.* [304] presented a learning-based method for constructing a driving signal recognition whole body avatar. They generate high-quality representations of human geometry and view-dependent shapes using conditionally deformable auto-encoders that are animated with imperfect driving signals (e.g., human poses and face key points). Better drivability and generalization were achieved by separating the unusable driving signals and the rest of the generated elements during animation.

Modeling thin structures (e.g., hair) has low resolution and is too slow. Lombardi *et al.* [305] showed a dynamic 3D content rendering representation that combines the completeness of a physical representation with the efficiency of primitive-based rendering. It utilizes spatially shared computations with a convolutional architecture and uses volumetric primitives that are moved to cover only the occupied portion of space. Sun *et al.* [306] introduced a hair inverse rendering framework for reconstructing high-fidelity 3D geometry and reflectivity of hair that is easily used for realistic rendering of hair. They proposed a new solution for line-based multi-view stereo that calculates accurate hair geometry from multi-view metering data and estimate hair reflection characteristics using multi-view metering data.

In the Metaverse, avatar clothing is not just for decoration but a means of providing immersion and emphasizing social roles. To create high-definition animations, Xiang *et al.* [307] proposed a method to create an animable clothed body avatar by explicitly representing the upper body's clothing in a multi-view capture video, as shown in Fig. 12. To separately register 3D scans with the template using a two-layer mesh representation and to improve photometric responsiveness, they perform texture alignment through the inverse rendering of the garment geometry and texture predicted by the deformation autoencoder. Chaudhuri *et al.* [308] proposed ReAVAE (Region-adaptive Adversarial Variational Variational AutoEncoder) that learns the probability distribution of each region individually to generate various high-fidelity texture maps for 3D human meshes by sampling from the distribution for each region. They present a data generation

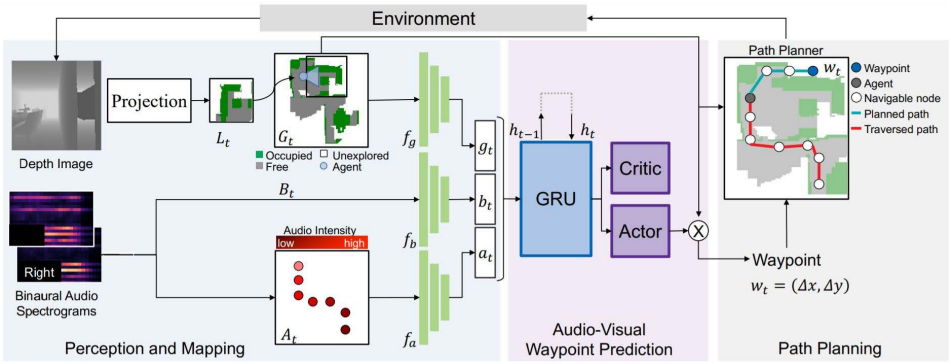

FIGURE 11. Learning to set waypoints for audio-visual navigation in the indoor environment [301].

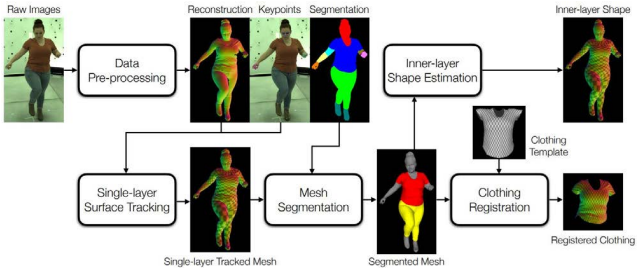

FIGURE 12. The process of cloth rendering which includes single-layer surface tracking and inner-layer shape estimation [307].

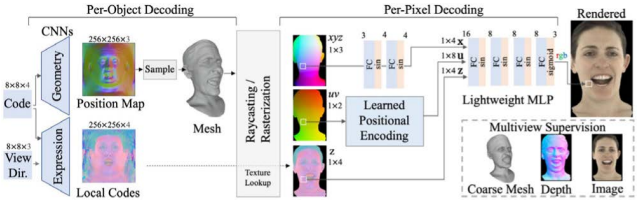

FIGURE 13. Face rendering with per-object decoding and per-pixel decoding in Pixel Codec Avatar [310].

technique that augments the training set with data taken from a single view RGB input.

It can be generalized to natural lighting conditions, but it is computationally expensive to render. Bi *et al.* [309] presented a method to build animable high-definition 3D face models that can pose and render in real-time in a novel lighting environment. They train a generalizable model and use it to generate a training set of high-quality synthetic face images under natural lighting conditions. The neural shading phase accounts for deformations that are not captured in the mesh and alignment inaccuracies and dynamics that confound the DNR pipeline. Raj *et al.* [207] proposed Articulated Neural Rendering (ANR), a DNR-based framework that explicitly addresses the limitations of virtual human avatars. Ma *et al.* [310] proposed Pixel Codec Avatars (PiCA), a deep generation model of the 3D human face that is computationally efficient and adaptable to in-run rendering conditions while achieving state-of-the-art reconstruction performance as depicted in Fig. 13. They use a fully convolutional architecture for decoding spatially varying features and a rendering adaptive per-pixel decoder to integrate through dense surface representations learned in a weakly supervised manner from low-topology mesh tracking on training images. It is strong at testing expressions and opinions about people of different genders and skin colors.

5) METAVERSE IMPLEMENTATIONS

a: MULTIMODAL INFERENCE

Self-supervised pre-training can outperform full-supervised training and is useful in preventing overfitting to smaller data sets. Shukla *et al.* [311] showed the potential of visual self-supervision for learning audio functions. They proposed that joint visual and audio self-supervision leads to more informative audio representations for speech and emotion recognition. The proposed multi-task combination of visual and auditory self-supervision is useful for learning more powerful and rich functions in noisy conditions.

b: RL-BASED APPROACHES

Procedurally generated environments require algorithmically generated environment instances using a unique variable factor configuration as an important benchmark for testing systematic generalization in deep reinforcement learning. Jiang *et al.* [312] proposed Prioritized Level Replay (PLR), a general framework for selectively sampling the next level of training by prioritizing items that are expected to have higher learning potential upon future revisit. TD errors lead to new curricula of increasingly difficult levels when used to effectively estimate the future learning potential of a level and guide the sampling procedure. Modhe *et al.* [313] proposed a novel framework that provides exploration and sample complexity to identify sub-objectives that are useful for exploration in sequential decision-making tasks under partial observability. They utilized a variant-specific control framework that maximizes empowerment to reach various states reliably. It identifies sub-goals as states with high essential, optional information through the normalization of information theory.

To efficiently control dynamic systems in high-dimensional sensory observations, learning controllable embeddings (LCEs) embed observations in low-dimensional latent space and estimate latent dynamics to perform control in latent space. Cui *et al.* [314] proposed a modified value-guided CARL that optimizes the weighted version of the

CARL loss function whose weights depend on the TD-error of the current policy. In the offline implementation, the local linear control algorithm (e.g., iLQR) used in the existing LCE method was replaced by the RL algorithm (i.e., a model-based soft actor-critic). Model-based reinforcement learning is a method that utilizes control-based domain knowledge to improve the sample efficiency of reinforcement learning agents. Policies tend to lag behind model-free agents in terms of final rewards, especially in environments where they are not critical. Amos *et al.* [315] found an effective combination of model-free soft value estimation for policy evaluation and model-based stochastic value gradient for policy improvement for model-based high-dimensional humanoid control tasks.

c: LIFE-LONG LEARNING

Sukhbaatar *et al.* [316] proposed Expire-Span, which learns how to retain the most important information and expire irrelevant information, as not all past contents need to be remembered equally. To evaluate models for lifelong learning tasks, Abdelsalam *et al.* [317] developed a standardized benchmark that enables model evaluation in IIRC settings. Methods incorporating network scaling naturally add model capacity for learning new tasks while avoiding catastrophic oblivion, but increasing the number of additional parameters is computationally expensive at larger scales. Verma *et al.* [318] proposed a simple task-specific feature map transformation strategy for continuous learning called Efficient Feature Transformations (EFT). It adds a minimal number of parameters to the underlying architecture, providing strong flexibility to learn new tasks. To solve the catastrophic forgetting problem in a sequential task where data from previous tasks are not available, Mehta *et al.* [319] proposed a principled Bayesian nonparametric approach, the Indian Buffet Process (IBP), which determines how much the data scales to model complexity. The IBP dictionary promotes positive knowledge transfer between tasks by encouraging sparse weighted element selection and element reuse. The goal of continuous learning (CL) is to learn a series of tasks without experiencing catastrophic forgetting. Ebrahimi *et al.* [320] proposed a simple educational paradigm, Remembering for Right Reasons (RRR), by encouraging explanations so that models have the right reasons for their predictions.

d: MULTI-AGENT OPTIMIZATION

The benefit of multi-task learning is that it uses relationships across tasks to improve the performance of a single task. Metadata is useful for improving multi-task learning performance, but effective integration is an additional challenge. Sodhani *et al.* [321] showed state-of-the-art results in Meta-World, which consists of a challenging multitasking benchmark. It learns expressions that are interpreted as metadata and helps provide context to tell which expressions to construct and how to construct them. Fu *et al.* [322] proposed a framework LeTS that utilizes multi-task computation and parameter sharing for efficient fine-tuning. It decouples the computational dependencies of existing fine-tuning models with a neural architecture that reuses intermediate results and reduces computational demands by leveraging the sparsity feature of weight differences. Zhang *et al.* [323] proposed Hidden-Parameter Markov Decision Processes (HiP-MDPs), an explicit modeling method for this structure, to improve sample efficiency in multitasking settings. In the HiP-MDP setting, they utilized the idea of a common structure and extended to enable state abstraction inspired by block MDP.

Dollar *et al.* [324] proposed a simple and fast complex scaling strategy that scales the underlying convolutional network to give greater computational complexity and, consequently, expressive power, extending the model. It provides a framework for analyzing scaling strategies under various computational constraints. Ruiz and Verbeek [325] proposed Hierarchical Neural Ensembles (HNE) to handle scenarios where the amount of computation and input data varies with time. It includes an ensemble of multiple networks in a hierarchical tree structure that shares an intermediate layer. As a hierarchical distillation to increase the prediction accuracy of small ensembles, the overlapping structure of the ensembles is utilized to allocate accuracy and diversity across individual models optimally.

e: INTEGRATION OPTIMIZATION

GPU performance and efficiency of recommendation models are affected by model architecture configurations (e.g., dense and sparse features and MLP dimensions). Acun *et al.* [326] described the complexity of using GPUs for training recommendation models, factors influencing hardware efficiency at scale, and a new scale-up GPU server design from Zion. Silent Data Corruption (SDC) is a negative impact on large-scale infrastructure services. Dixit *et al.* [327] provided a debug flow based on the root cause and classification error guidance within the CPU using case studies as an explanation of how to debug this class of errors.

Vanilla NAS provides real-world performance, as each architecture is evaluated through training from scratch, but it is time-consuming. Zhao *et al.* [328] showed that one-shot NAS significantly reduces the computational cost by training only one supernetwork to approximate the performance of all architectures in the search space through weight sharing. To mitigate unwanted joint adaptation, they proposed several NAS using multiple supernetworks, called sub-supernets, each covering different areas of the search space. Stage 2 NAS needs to sample from the search space during training, which directly affects the accuracy of the final searched model. Uniform sampling has been widely used for simplicity but is agnostic to the model performance Pareto front, which is the primary focus of the search process, thus missing the opportunity to improve model accuracy further. Wang *et al.* [329] proposed an AttentiveNAS that focuses on enhancing the sampling strategy.

MBRL algorithms are complex due to the separate dynamic modeling and follow-up planning algorithms.

Consequently, when possessing dozens of hyperparameters and architectural choices, significant human expertise is required before applying them to new problems and domains. Zhang *et al.* [330] used automatic hyperparameter optimization (HPO) to improve performance compared to using static hyperparameters fixed for the entire training during training itself. Zhang *et al.* [331] studied how expressive learning can accelerate reinforcement learning from rich observations (e.g., images) without relying on domain knowledge. The bi-simulation metric quantifies the behavioral similarity between states in continuous MDP and trains the encoder so that the distance in the latent space is equal to the bi-simulation distance in the state space. For robust, fast, and scalable binary optimization, Panchenko *et al.* [332] proposed Lightning BOLT, an improved version of the BOLT binary optimizer that significantly reduces the processing time and memory requirements while maintaining the efficiency of the BOLT, which enhances the performance of the final binary.

f: OPERATION CONSIDERATION

Empirical risk minimization (ERM) is generally designed to perform well for mean loss so that the estimator is sensitive to outliers, does not generalize, and treats subgroups unfairly. Li *et al.* [333] explored the problem through an integrated framework called TERM (Tilted Empirical Risk Minimization) to increase or decrease the impact of outliers, respectively. They show that TERM is used in a variety of applications (e.g., enhancing fairness between subgroups, mitigating the effect of outliers, and handling class imbalance).

Fairness and robustness are two important concerns of federated learning systems. Robustness to data and model poisoning attack and fairness are the constraint to compete in statistically heterogeneous networks, as measured by the uniformity of performance across devices. Li *et al.* [334] proposed to use Ditto, a simple and general framework for personalized federated learning and developed an extensible solver for this. To understand and improve fault tolerance training for deep learning recommendations with partial recovery, Maeng *et al.* [335] optimized CPR, a partial recovery training system for a recommendation model. It relaxes consistency requirements and improves failure-related overhead.

Unexpected reboots disrupt services running on the hardware and reduce fleet availability. A server reboot is also an important signal that indicates an underlying problem (e.g., a memory leak in service, a catastrophic hardware failure, a power outage) in a data center. Lin *et al.* [336] provided a large-scale, near-real-time reboot-monitoring framework that supports machine learning-based anomaly detection and automated root cause analysis for hundreds of server attribute combinations. Xia *et al.* [337] proposed Facebook's risk-focused backbone management strategy to ensure high service performance during the COVID-19 pandemic. It has been shown to achieve high service availability and low path scalability while resiliently withstand stress tests and handles traffic spikes efficiently.

Oughton *et al.* [338] anticipated that 5G would remain the preferred technology for wide-area coverage, while Wi-Fi 6 will remain the preferred technology indoors thanks to its much lower deployment costs. To address the problem of packet loss affecting a wide range of applications using Voice over IP (VoIP), Lin *et al.* [339] proposed prediction and mask training to improve the performance of the CRN framework. It outperforms the reference system using only the LSTM layer in terms of two objective metrics: speech quality (PESQ) and short-term objective intelligibility (STOI). The CRN consists of a convolutional encoder-decoder structure and an LSTM (long short-term memory) layer that is suitable for real-time speech enhancement applications.

Applying homogeneous encryption (HE) to the client-cloud model allows the cloud service to perform inference directly on the client's encrypted data. However, HE satisfies privacy constraints, but it introduces enormous computational problems in the current system. Reagen *et al.* [340] introduced Cheetah, a set of algorithms and hardware optimizations for server-side HE DNN inference to approach real-time speed. Automatic compilation of an efficient HE kernel in a synthetic compiler for vectorized isomorphic encryption is relatively unexplored. Cowan *et al.* [341] proposed an optimizing compiler, Porcupine, which uses program synthesis to generate vectorized HE code. Porcupine captures the underlying HE operator behavior and automatically infers the complex trade-offs imposed by these issues to develop an optimized and validated HE kernel.

6) METAVERSE APPLICATIONS

a: SIMULATION

Carrying suspended payloads is difficult for autonomous aircraft, and rapid in-flight adaptation to payloads with physical properties unknown a priori remains an open question. Belkhale *et al.* [342] proposed a meta-learning approach that learns to learn a modified dynamics model within seconds of flight data after connection. One way to infer the safety of a robot is to build a safe set through Hamilton-J, but because of the long computation time, it sometimes assumes perfect knowledge of the mechanics, and the safety set is calculated offline. Shih *et al.* [343] proposed a new framework for learning safety control policies from simulation and using it to generate online safety sets from uncertain dynamics. As climate change increases the frequency and severity of natural disasters, response organizations need improved data to better understand the dynamics of disaster impacts. Giraudy *et al.* [344] are leveraging Facebook Location History (LH) data as part of its disaster mapping initiative to enable location-based services (e.g., Nearby Friends, location-based advertising) and social value products (e.g., disaster maps) to help people locate.

b: GAME

Diplomacy is a game of switching alliances that involves both cooperation and competition, which is not successful in large-scale games involving collaboration. Gray *et al.* [345] described a media-free Diplomacy transforming agent that combines supervised learning on human data with one-step preview search through minimizing external regrets.

c: OFFICE

Ha-Thuc *et al.* [346] discussed how these systems evolve from traditional formulations by incorporating producer values into goals. Jointly optimizing the ranking function for both consumer and producer value is a new direction and raises many technical challenges. They make the layout an end-to-end solution and describe the results of applying it to Facebook Marketplace. Blackshear *et al.* [347] proposed a method for blockchain asset owners to recover their funds if their private key is accidentally lost or sent to the wrong address. They achieve this with a Commit, Reveal, Claim, and Challenge smart contract that allows access to funds at addresses where the spend key is unavailable. The auction market introduces the concept of speed balance by reinterpreting the process of applying a coefficient between 0 and 1, which equalizes bids in all auctions on behalf of each buyer. Conitzer *et al.* [348] showed that calculating the social welfare maximization and profit maximization rate equilibrium is NP-hard but presents a mixed-integer program (MIP) is used to find a balance that optimizes several related goals. It uses static MIP solutions to improve the results achieved with dynamic pacing algorithms using instances based on real auction markets.

d: SOCIAL

Online social network (OSN) accounts exhibit many demographic attributes (e.g., age, gender, location, and occupation). Onaolapo *et al.* [349] devised a method to instrument and monitor stolen social accounts to understand the impact of demographic characteristics on attacker behavior. Cybercriminals accessing teen accounts create more messages and posts than cybercriminals accessing adult accounts, while attackers compromising male accounts destroy, including changing some of their profile information. Cybercriminals accessing female accounts appeared to be engaged in hostile activity. Bailey *et al.* [350] explored the spatial structure of social networks in the New York metropolitan area, where a significant proportion of city dweller connections are with nearby individuals. By examining the importance of transport infrastructure, they document significant heterogeneity in the geographic extent of social networks and show that this heterogeneity is correlated with public transport use. In the present state of sharing both temporary and permanent content on social media platforms, Luria and Foulds [351] discussed our findings on the short-term and long-term transitivity as part of social media experiences and the evolving identities of teens and young adults. As long as proportionality is not violated, there are greedy algorithms that involve volunteers and non-adaptive methods that include volunteers with trait-only probabilities assuming that the distribution of common traits in the volunteer pool is known. Although this distribution is not known a priori, Do *et al.* [352] proposed a reinforcement learning-based approach for online learning.

e: MARKETING

Fernanda [353] proposed the knowledge framework by using a mix of quantitative and qualitative methods to explore the current state of diversity and representation in online advertising and people's attitudes to the impact of diversity on digital campaign performance. More frequent and positive portrayals of underrepresented and diverse groups have a significant positive impact on business outcomes. It is important to optimize advertisers' budgets for campaigns across platforms without knowing the value of serving ads to users on multiple platforms. Avadhanula *et al.* [354] provided a regret algorithm for individual bid spaces. The generalization of existing MAB algorithms (e.g., Upper Confidence Bound and Thompson Sampling) does not perform well in two applications: the intelligent SMS routing problem and the advertising audience optimization problem that many businesses (especially online platforms) face. Sinha *et al.* [355] presented a simple variant of explore-the-commit and improved performance by setting a near-optimal regret range for this algorithm.

f: EDUCATION

Because simulation provides the ability to train a large number of robots in parallel and provides rich data, Truong *et al.* [356] used educational simulations before deploying the robots. They proposed bidirectional domain adaptation (BDA), an approach that connects the sim-vs-real gap in both directions for point goal navigation. They use Real2sim for bridging the visual domain gap and sim2real for linking the dynamic domain gap.

7) DISCUSSION AND OPEN CHALLENGES

As mentioned above, Facebook research is a research group with a lot of interest in Metaverse, as shown in Table 2. It has broad elemental technologies for natural language, vision, dialogue, and embodiment. It also has a foundation and experience that is expanded into a Metaverse platform with a Facebook social network service. Essential models for Metaverse are Blenderbot [286] based on PariAI, Detectron 2 [357] capable of fast visual recognition, and Habitat [300] that operate an agent from an eco-centric point of view. They provide services by launching its own Metaverse platform, Horizon and Infinite Office. The virtual currency Dime not only serves as a bridge between reality and the Metaverse but also leads to a sustainable ecosystem.

VI. DISCUSSION AND OPEN CHALLENGES

In this section, we discuss current problems and technologies needed in the future for Metaverse in the aspect of influence, limitations, and open challenges.

TABLE 2. Use case - Facebook Metaverse.

Vendor	Paper	Method
Physical devices and sensors	Head-Mounted Displays (HMD)	Li et al. [23], Xian et al. [234], Bonatti et al. [235]
	Hand-based input device	Foy et al. [74], Zhang et al. [236], Rognon et al. [237]
	Non-hand-based input device	Chase et al. [238], Stephens-Fripp et al. [239]
	Motion input device	Yuan et al. [240], Neverova, et al. [241]
Recognition and rendering	Scene and object recognition	d'Ascoli et al. [243], Porzi et al. [244], -Video (Jin et al. [246], Xiong et al. [247]), -OCR (Huang et al. [248], Singh et al. [249]), -3D (Shagun et al. [250], Henzler et al. [251], Ren et al. [75]), -SSL (Liu et al. [252], Xinlei and He [253], Yuandong et al. [254])
	Sound and speech recognition	-Multi-source (Donley et al. [255], Tzirakis et al. [256]), -Noise (Helmholz et al. [257], Chazan et al. [258]), -DoA (Shlomo and Boaz [259], Shlomo and Boaz [260]). -HRTF (Zhou et al. [261], Ben-Hur et al. [262], Chojnacki et al. [263], Gari et al. [264])
	Scene and object generation	Ge et al. [265], Roziere et al. [266], Lassner and Zollhofer [267],
	Sound and speech synthesis	Wu et al. [268], Richard et al. [269]
	Motion rendering	Won et al. [270], Ye et al. [271], Yuan et al. [272], Ng et al. [273], Eisenberger et al. [274]
Scenario generation	Multimodal content representation	Chen et al. [275]
	Agent persona modelling	Kiela et al. [276], Blevins et al. [277]
	Multimodal entity linking and expansion	De Cao et al. [278]
	Scenario generation	Gafni et al. [279]
	Scenario population	Gong et al. [280], Li et al. [281], Liang et al. [282]
	Scenario evaluation	Jia et al. [283], Huang et al. [284], Ahlgren et al. [285]
User Interaction	Language interaction	Dagan et al. [287], -LM (Le et al. [288], Weber et al. [289]), -QA (Lewis et al. [290], Xiong et al. [291] , Min et al. [292]) -Translation (Schwenk et al. [293], Tang et al. [294], Tuan et al. [295])
	Multimodal interaction	-Video (Patrick et al. [296], Meng et al. [297]), -Audio (Morgado et al. [298], Morgado et al. [299])
	Multi-task interaction	Szot et al. [300], Chen et al. [301], Chen et al. [302], Ye et al. [303]
	Embodied interaction	-Body (Bagautdinov et al. [304], Lombardi et al. [305]), -Avartar (Xiang et al. [307], Chaudhuri et al. [308], Bi et al. [309], Ma et al. [310])
Implementations	Multimodal inference	Shukla et al. [311]
	RL-based approaches	Jiang et al. [312], Modhe et al. [313], Cui et al. [314], Amos et al. [315]
	Life-long learning	Sukhbaatar et al. [316], Abdelsalam et al. [317], Verma et al. [318], Ebrahimi et al. [320]
	Multi-agent optimization	Sodhani et al. [321], Fu et al. [322], Zhang et al. [323], Dollár et al. [324], Ruiz and Verbeek [325]
	Integration optimization	-HW (Acun et al. [326], Dixit et al. [327]), -NAS (Zhao et al. [328], Wang et al. [329]), -HPO (Zhang et al. [330], Zhang et al. [331], Panchenko et al. [332])
	Operation consideration	Li et al. [333] -Robustness (Li et al. [334], Maeng et al. [335]), -Backup (Lin et al. [336], Xia et al. [337]), -Network (Oughton et al. [338], Lin et al. [339]), -Security (Reagen et al. [340], Cowan et al. [341])
Metaverse applications	Simulation	Belkhale et al. [342], Shih et al. [343], Giraudy et al. [344]
	Game	Gray et al. [345]
	Office	Ha-Thuc et al. [346], Blackshear et al. [347], Conitzer et al. [348]
	Social	Onaolapo et al. [349], Bailey et al. [350], Luria and Foulds [351], Do et al. [352]
	Marketing	Fernanda [353], Avadhanula et al. [354], Sinha et al. [355]
	Education	Truong et al. [356]

A. METAVERSE INFLUENCE FOR USER AND SOCIETY

1) SENTIMENT AND SOCIAL INFLUENCE

People can lead a stable cyber life in the Metaverse because they can distinguish between real-life and virtual life, just as a person did not feel confused while watching an avatar movie. However, because avatar design has emotional barriers, users may feel a sense of rejection towards the avatar if it cannot overcome Uncanny Valley like in the Alita movie.

Memories are beautiful because they are exquisitely crafted memories but because they are traces of time that cannot be returned. However, Metaverse recreates the past, and users can make different choices, giving people psychological stability and emotional recovery.

While the social impact depends on the ecosystem, it is important to consider many aspects of social impact, including potential exacerbation of social inequities, computation demands, economics, legality, and ethics. In addition, limited resources in the real world bring excessive competition and social side effects. However, in the Metaverse, it has an advantage over the real-world system in terms of item production and resources. It is possible to use infinite resources that can be created indefinitely online rather than a deduction compensation from limited resources in the real-life world. This is different from the reward system in the real world, where you must give up the other to get one, so it can reduce competition between users and is an opportunity to develop for the common good.

2) USER PARTICIPATION AND BENEFIT

In the Metaverse, users are less limited by time and space and can exist in multiple places through avatars, so the communication style is changed from 1:N broadcasting to 1:1 interaction. Avatar in the Metaverse provides a way to replace and complement users. The Metaverse is most effective in places like Africa where experiential education is difficult (e.g., undeveloped areas). In addition, High-contrast vision, long-distance vision, and volume augmentation for people with visual difficulties enable people with disabilities to live the same lives as ordinary people in Metaverse.

Mask effects (i.e., hide shapes, colors, and races) are also noteworthy, providing a better-than-realistic user engagement experience. Origin, gender, skin color, and appearance can be prejudicial when it comes to debates, psychiatric group therapy, and jury attendance in court. In this case, the avatar's neutral appearance is a good example of the Metaverse's social influence, which allows for a fairer opinion and participation in social consensus without prejudice.

In order to maintain a sustainable social ecosystem, user participation is important, so it is essential to provide fashion, games, and events on a regular and long-term basis. For example, in ZEPETO, users take selfies, solve quizzes, create dramas, and have fun designing costumes. In Metaverse sports and gun games, it is possible to induce and increase user participation by providing a third-person view rather than a first-person spectator mode.

3) MORE APPLICATIONS

Metaverse can significantly contribute to a multitude of applications and domains. However, for sustainable Metaverse application, we must consider interpretability, security, privacy, societal function, and ethics. More Metaverse applications help people work smart (e.g., telemedicine, layer separation and tagging for complex organ surgery, commuting to work simulation, remote problem identification, mapping to real environments without the need to find manuals). Metaverse makes living easy (e.g., senior public transportation simulation, intuitive simplification of digital input interfaces for the elderly, immersive education more effective than a video, counseling personal issues with masked avatar). Metaverse applications reduce physical object and space (e.g., non-shared private messaging in the desired place, providing information through overlay display of offline objects, virtual screen, store inventory trends, sales volume display, virtual display for IoT devices, and smart home applications).

B. METAVERSE LIMITATION

1) SUSTAINABILITY

Many advantages and applications have been described, but the sustainability of Metaverse is important. When the world's population is maintained at a certain level, it can grow and fix problems, but when the number of users accessing decreases, the world cannot be maintained. In the concept of life logging, the sustainability of various social relationships is more important than each event and task (e.g., games and simulations). In order to maintain continuity, a connection relationship (e.g., Metaverse access, messenger) must be maintained continuously in a relatively low-spec mobile device that can always be accessed. Using an episodic memory that effectively manages the user's log allows the user to feel the comfort and advantage of accessing Metaverse for a long time. Storing all experiences in memory storage has limitations in utilization and capacity, so memory research on effectively finding and reusing important episodes is needed. In addition, latecomer platforms should consider import/export methods that bring the existing user experience and provide continual usability.

2) HARDWARE AND SOFTWARE LIMITATIONS

In terms of a sensor in hardware, while the Metaverse resembles the real world a lot, some sensations are better felt in real life (e.g., day sunlight, smell, stickiness, slippery, wind). In terms of software, programs developed in the Metaverse without coding are used as a basis for high compatibility in the Metaverse world. However, as the program becomes more sophisticated, it faces the limit of sophistication in a complex application.

In terms of content, the dialog is developing into a longer and more natural form of conversation based on persona, but it is still limited as a sustainable lifelong-learning conversation solution with various perspectives and philosophies beyond exciting conversations. Humans basically have multi-personas, and they are expressed differently depending on time and place. Therefore, it is necessary to study more complex persona modeling in consideration of the situation. From this point of view, environments and events are important to show the various personas of users and NPCs. For example, in the drama Westworld, avatars

perform various actions in the Metaverse, freeing from the constraints and conditions of reality. Therefore, NPCs in the role of residents of the Metaverse must be able to cope with various unexpected situations because the allowable range of scenarios is wide. In addition, the persona's design is important for the NPC to appear as if they choose with their persona and will. NPCs can be in the form of humans and various living (e.g., horse, dog, cat) and non-living forms (e.g., desk, clock).

3) DEVELOPMENT HUDDLE

In Metaverse, since it is a comprehensive solution in which various tasks occur simultaneously in a complex form (e.g., multi-mode and multi-task), there is a lot of work to study for individuals to start development without experience. From the perspective of Metaverse development, there are few online resources to learn, especially for novice developers. There is not enough information for practical details to make complex and realistic implementations (e.g., object selection, conditional actions, user storyboards with scene flow, teleportation between scenes, movement, and dialogue). For this reason, a collaborative system (i.e., a platform and developer community) for an individual developer is important to co-develop without designing the entire system. As for the platform, a commercial platform (e.g., Roblox) with favorable maintenance and an open source-based platform (e.g., Unity) with various possibilities are considered. Since the scope of the technology target of Metaverse is wide, it is necessary for the developer community to separate threads based on well-organized taxonomy and maintain a group of experts who lead in each technical domain.

C. OPEN CHALLENGES

1) MEDIUM SELECTION

AR uses lightweight devices, suitable for short experiences, but VR relatively needs heavy and expensive devices for long experiences. Some approaches switch between AR and VR in one piece of hardware by mixing the advantages of AR and VR. Although this method has the advantage of using AR and VR in alternative ways, it becomes expensive and heavy compared to a single model device. Alternatively, holograms are not a popular technology in Metaverse, but they have potential.

Eye-worn lenses are another input method utilized in the Metaverse (e.g., Maya Lenz, Mirage, Mojo lenz). The lens analyzes the user's information by tracking the direction of eye movements, focus, blinks, and winks. For example, Maya Lenz is a wearable device in the form of a contact lens, and Mirage is a way of expressing disliked content by replacing it with positive alternatives. Mojo lenz is used in conjunction with an assistive device worn around the neck to seamlessly process a variety of visual information (e.g., data feeds, people's profiles, video calls, translations, notifications) into the wearer's vision.

2) ETHICS AND SECURITY

Privacy and security are critical issues because Metaverse collects data on behavior that is more detailed than user conversations and internet history. Avatar two-factor authentication and protection of transmitted data are essential, and we need to be more vigilant with regard to crimes that may occur on the Metaverse. In addition, surveillance actions (e.g., inappropriate chat room surveillance, censorship, and follow-up review) due to the surge in users suggest that organizations that play the same role as police and government are needed in the real world. There are some instances where exemplary people in the real world commit crimes based on their online anonymity in the Metaverse. The norms and restrictions of the Metaverse may differ from those in the real world because they have a post-nationalism and degrees of freedom. Most users familiar with the Metaverse are the young generation with relatively various social ideas. It is necessary to build a Metaverse with a worldview and ethical consciousness in which various avatars can live, rather than a Metaverse as a physical space.

3) INTERDISCIPLINARY RESEARCHES

Since the Metaverse consists of a world that changes in real-time for a large number of users and NPCs, cross-disciplinary research is necessary. As an example of cross-disciplinary research, Metaverse leverages knowledge widely used in cognitive science (e.g., episodic memory, intrinsic motivation, and theory of mind) to provide more immersive and sustainable services. Episodic memory occurred a long time ago in the present conversation and induced a natural conversation. Intrinsic motivation allows an agent to perform multiple tasks rather than a single task consistently. The theory of mind has the advantage of deepening conversation to understand from the other person's point of view.

Other examples are the social sciences, psychology, and economics. The environment in which a certain number of members live using masked avatars differs from how society currently operates. Neuroscience and psychological approaches for psychotherapy are used to understand humans and maintain a Metaverse deeply. The virtual currency of Metaverse is different from the virtual currency in the real world in that it is used as a real product in a virtual environment, so it can become a new variable from the point of view of economics and develop into a fused form.

VII. CONCLUSION

In this study, we analyzed research for similar concepts of Metaverse in Metaverse, avatar, and XR. After that, we comprehensively dealt with the necessary three components (i.e., hardware, software, and contents) for Metaverse. We also reviewed the latest trends of Metaverse approaches (i.e., user interaction, implementation, and application) that were currently available and necessary in the future. Interacting as part of the story is important rather than seeing well-formed storytelling and immersive visual effects. We applied

taxonomy to three famous Metaverse domains (i.e., movie, game, and researches) in Ready Player One, Roblox, Facebook Research. Finally, we discussed the aspect of social influence, limitation, and open challenges.

From a future-oriented perspective, Facebook research tries to input text using the output of the peripheral nervous system and brain-computer interface. As a direct connection method, Neuralink is a way to enhance communication with devices by implanting a chip in the human brain. The current stage of development is to the extent that it is possible to directly stimulate a specific part of the brain and look at a simple type of EEG. However, the continuous development of brain-computer-interface and Neuralink can develop into a form that gives an experience that is difficult to distinguish from reality in the Metaverse (e.g., the method of connecting to the spine from the matrix).

APPENDIX

TABLE 3. List of main acronyms.

Acronym	Full Form	Acronym	Full Form
AGI	Artificial General Intelligence,	**MLRF**	Multimodal Local-Global Ranking Fusion
ANR	Articulated Neural Rendering	**MMORPGs**	Multiplayer Online Role-Playing Games
AP	Average Precision	**MONET**	Multi-Object Network
AR	Augmented Reality	**NMT**	Neural Machine Translation
BDA	Bidirectional Domain Adaptation	**NPC**	Non-Player Character
BERT	Bidirectional Encoder Representations from Transformers	**NPIs**	Negative Polarity Items
CABB	Crop Recognition Bounding Box Regression Loss	**OCR**	Optical Character Recognition
CGI	Computer-Generated Imagery	**OSN**	Online Social Network
CL	Continuous Learning	**PBL**	Problem-Based Learning
CNN	Convolutional Neural Network	**PHALCOR**	PHase ALigned CORrelation
CURL	Contrastive Unsupervised Representations for Reinforcement Learning	**PiCA**	Pixel Codec Avatars
DIAYN	DIversity All You Need	**PLMs**	Pre-trained Language Models
DOA	Direction Of Arrival	**PLR**	Prioritized Level Replay
ECL	Natural language in Cognitive Linguistics	**PQ**	Panoptic Quality
EFT	Efficient Feature Transformations	**QACNN**	Query-based Attention CNN
EMIL	Embodied Multi-modal Interaction in Language learning	**ReAVAE**	Region-adaptive Adversarial Variational Variational AutoEncoder
EQA	Embodied Question and Answer	**ReTiSAR**	Real-Time Spherical Array Renderer
ERM	Empirical Risk Minimization	**RL**	Reinforcement Learning
FED	Foveated Entropic Differencing	**RNN-T**	Recurrent Neural Network Transducer
FR	Full Reference	**RRR**	Remembering for Right Reasons
GAN	Generative Adversarial Network	**SDC**	Silent Data Corruption
GCN	Graph Convolution Network	**SDM**	Spatial Decomposition Method
GNN	Graph Neural Network	**SIR**	Spliced Image Retrieval
GPSA	Gated Positional Self-Attention	**SLU**	Spoken Language Understanding
HE	Homogeneous Encryption	**SQ**	Speech Quality
HIMN	Hierarchical Interactive Memory Network	**SS-OD**	Semi-Supervised Object Detection
HiP-MDPs	Hidden-Parameter Markov Decision Processes	**STOI**	Short-Term Objective Intelligibility
HMD	Head Mounted Display	**TERM**	Tilted Empirical Risk Minimization
HRTF	Head-Related Transfer Function	**ToM**	Theory of Mind
I2A	Imagination-Augmented Agents	**VAE**	Variational AutoEncoder
IBP	Indian Buffet Process	**ViTs**	Vision Transformers
KD	Knowledge Distillation	**VLN**	Visual Language Navigation
KG-A2C	Knowledge Graph A2C	**VLP**	Unified Vision Language Dictionary Training
LBM	Local Bipartite Matching	**VoIP**	Voice over IP
LCEs	Learning Controllable Embeddings	**VPMs**	Video Pre-trained Models
LCMV	Linearly Constrained Minimum Variance	**VQA**	Visual Question Answering
LH	Location History	**VR**	Virtual Reality
LSTM	Long Short-Term Memory		

REFERENCES

[1] *Roblox Blog*, Roblox Corp., 2022. Accessed: Nov. 1, 2021. [Online]. Available: https://blog.roblox.com/

[2] C. Meier, J. Saorín, A. B. de León, and A. G. Cobos, "Using the Roblox Video Game Engine for Creating Virtual tours and Learning about the Sculptural Heritage," *Int. J. Emerg. Technol. Learn.*, vol. 20, pp. 268–280, Oct. 2020.

[3] R. U. Long, "Roblox and effect on education," M.S. thesis, Dept. Educ., Drury Univ., Springfield, MO, USA, 2019.

[4] J. Kemp and D. Livingstone, "Putting a second life 'metaverse' skin on learning management systems," in *Proc. 2nd Life Educ. Workshop Second Life Community Conv.*, San Francisco, CA, USA, vol. 20, 2006, pp. 22–47.

[5] A. M. Kaplan and M. Haenlein, "The fairyland of second life: Virtual social worlds and how to use them," *Bus. Horizons*, vol. 52, no. 6, pp. 563–572, Nov. 2009.

[6] M. R. Cagnina and M. Poian, "How to compete in the metaverse: The business models in second life," U Udine Econ. Work. Paper, Udine, Italy, Tech. Rep. 01-2007, 2007, doi: 10.2139/ssrn.1088779.

[7] H. Duan, J. Li, S. Fan, Z. Lin, X. Wu, and W. Cai, "Metaverse for social good: A university campus prototype," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021 pp. 153–161.

[8] H.-S. Choi and S.-H. Kim, "A content service deployment plan for metaverse museum exhibitions—Centering on the combination of beacons and HMDs," *Int. J. Inf. Manage.*, vol. 37, no. 1, pp. 1519–1527, Feb. 2017.

[9] S.-N. Suzuki, H. Kanematsu, D. M. Barry, N. Ogawa, K. Yajima, K. T. Nakahira, T. Shirai, M. Kawaguchi, T. Kobayashi, and M. Yoshitake, "Virtual experiments in metaverse and their applications to collaborative projects: The framework and its significance," *Proc. Comput. Sci.*, vol. 176, pp. 2125–2132, Jan. 2020.

[10] B. Ryskeldiev, Y. Ochiai, M. Cohen, and J. Herder, "Distributed metaverse: Creating decentralized blockchain-based model for peer-to-peer sharing of virtual spaces for mixed reality applications," in *Proc. 9th Augmented Hum. Int. Conf.*, Feb. 2018, pp. 1–3.

[11] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering—A systematic literature review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, 2009.

[12] N. Stephenson, *Snowcrash*. London, U.K.: ROC, 1992.

[13] R. Schroeder, H. Avon Huxor, and S. Andy, "Activeworlds: Geography and social interaction in virtual reality," *Futures*, vol. 33, no. 7, pp. 569–587, 2001.

[14] C. Jaynes, W. B. Seales, K. Calvert, Z. Fei, and J. Griffioen, "The metaverse: A networked collection of inexpensive, self-configuring, immersive environments," in *Proc. Workshop Virtual Environ. (EGVE)*, 2003, pp. 115–124.

[15] C. Ondrejka, "Escaping the gilded cage: User created content and building the metaverse," *NYL Sch. L. Rev.*, vol. 49, p. 81, May 2004.

[16] B. Goertzel, "Human-level artificial general intelligence and the possibility of a technological singularity: A reaction to Ray Kurzweil's the singularity is near, and McDermott's critique of Kurzweil," *Artif. Intell.*, vol. 171, no. 18, pp. 1161–1173, 2007.

[17] C. Collins, "Looking to the future: Higher education in the Metaverse," *Educause Rev.*, vol. 43, no. 5, pp. 51–63, 2008.

[18] M. Wright, H. Ekeus, R. Coyne, J. Stewart, P. Travlou, and R. Williams, "Augmented duality: Overlapping a metaverse with the real world," in *Proc. Int. Conf. Adv. Comput. Entertainment Technol. (ACE)*, 2008, pp. 263–266.

[19] E. Schlemmer, T. D. Trein, and O. Cristoffer, "The metaverse: Telepresence in 3D avatar-driven digital-virtual worlds," *Tic. Revista d'innovaci? Educativa*, vol. 2, pp. 26–32, Jul. 2009.

[20] F. M. Schaf, D. Müller, F. W. Bruns, C. E. Pereira, and H.-H. Erbe, "Collaborative learning and engineering workspaces," *Annu. Rev. Control*, vol. 33, no. 2, pp. 246–252, Dec. 2009.

[21] G. Prisco, "A virtual world space agency," *Futures*, vol. 41, no. 8, pp. 569–571, Oct. 2009.

[22] M. Rymaszewski, W. J. Au, M. Wallace, C. Winters, C. Ondrejka, and B. Batstone-Cunningham, *Second Life: The Official Guide*. Hoboken, NJ, USA: Wiley, 2007.

[23] P. R. Messinger, E. Stroulia, K. Lyons, M. Bone, R. H. Niu, K. Smirnov, and S. Perelgut, "Virtual worlds—Past, present, and future: New directions in social computing," *Decis. Support Syst.*, vol. 47, no. 3, pp. 204–228, Jun. 2009.

[24] S. Hazan, "Musing the metaverse," in *Heritage in the Digital Era*. Brentwood, Esse, U.K.: Multi-Science Publishing, 2010.

[25] S. Papagiannidis and M. Bourlakis, "Staging the new retail drama: At a metaverse near you!" *J. Virtual Worlds Res.*, vol. 2, no. 5, pp. 425–446, Feb. 2010.

[26] M. Forte, N. Lercari, F. Galeazzi, and D. Borra, "Metaverse communities and archaeology: The case of Teramo," in *Proc. EuroMed*, Nov. 2010, pp. 79–84.

[27] T. C. Cunningham, "Marching toward the metaverse; strategic communication through the new media," Army Command Gen. Staff Coll Fort Leavenworth KS School Adv. Mil. Stud., VA, USA, Tech. Rep. ADA522953, 2010.

[28] D. Owens, A. Mitchell, D. Khazanchi, and I. Zigurs, "An empirical investigation of virtual world projects and metaverse technology capabilities," *ACM SIGMIS Database, Database Adv. Inf. Syst.*, vol. 42, no. 1, pp. 74–101, 2011.

[29] C. N. Tonéis, "Puzzles as a creative form of play in metaverse," *J. Virtual Worlds Res.*, vol. 4, no. 1, Jul. 2011.

[30] J. Guo, C. Angelina, and W. T. Rolf, "Virtual wealth protection through virtual money exchange," *Electron. Commerce Res. Appl.*, vol. 10, no. 3, pp. 313–330, 2011.

[31] T. M. Connolly, M. Stansfield, and T. Hainey, "An alternate reality game for language learning: ARGuing for multilingual motivation," *Comput. Educ.*, vol. 57, no. 1, pp. 1389–1415, Aug. 2011.

[32] A. Resmini and R. Luca, "*Pervasive Information Architecture: Designing Cross-Channel User Experiences*, vol. 4. Amsterdam, The Netherlands: Elsevier, 2011.

[33] F. Müller, "Remembering in the metaverse: Preservation, evaluation, and perception," Ph.D. dissertation, Dept. Mathematik Informatik, Univ. Basel, Basel, Switzerland, 2012.

[34] D. Xanthopoulou and S. Papagiannidis, "Play online, work better? Examining the spillover of active learning and transformational leadership," *Technol. Forecasting Social Change*, vol. 79, no. 7, pp. 1328–1339, Sep. 2012.

[35] A. Cameron, "Splendid isolation: 'Philosopher's islands' and the reimagination of space," *Geoforum*, vol. 43, no. 4, pp. 741–749, Jun. 2012.

[36] I. Hughes, "Virtual worlds, augmented reality, blended reality," *Comput. Netw.*, vol. 56, no. 18, pp. 3879–3885, Dec. 2012.

[37] C. Kim, S.-G. Lee, and M. Kang, "I became an attractive person in the virtual world: Users' identification with virtual communities and avatars," *Comput. Hum. Behav.*, vol. 28, no. 5, pp. 1663–1669, Sep. 2012.

[38] H. Kanematsu, T. Kobayashi, N. Ogawa, D. M. Barry, Y. Fukumura, and H. Nagai, "Eco car project for Japan students as a virtual PBL class," *Proc. Comput. Sci.*, vol. 22, pp. 828–835, Jan. 2013.

[39] G. Kipper and R. Joseph, *Augmented Reality: An Emerging Technologies Guide to AR*. Amsterdam, The Netherlands: Elsevier, 2012, ch. 1.

[40] S.-K. Kim, Y. S. Joo, M. Shin, S. Han, and J.-J. Han, "Virtual world control system using sensed information and adaptation engine," *Signal Process., Image Commun.*, vol. 28, no. 2, pp. 87–96, Feb. 2013.

[41] M. Preda, F. Morán, and C. Timmerer, "Introduction to the special issue on MPEG-V," *Signal Process., Image Commun.*, vol. 28, no. 2, pp. 85–86, Feb. 2013.

[42] A. Luse, B. Mennecke, and J. Triplett, "The changing nature of user attitudes toward virtual world technology: A longitudinal study," *Comput. Hum. Behav.*, vol. 29, no. 3, pp. 1122–1132, May 2013.

[43] J. D. N. Dionisio, W. G. B. Lii, and R. Gilbert, "3D virtual worlds and the metaverse: Current status and future possibilities," *ACM Comput. Surv.*, vol. 45, no. 3, pp. 1–38, Jun. 2013.

[44] E. Ko and J. Jang, "The virtual device managing module of the metaverse assisted living support system," in *Proc. Int. Conf. Modeling, Simulation Vis. Methods (MSV) Steering Committee World Congr. Comput., Eng. Appl. Comput. (WorldComp)*, 2014, pp. 125–126.

[45] M.-I. Dascalu, A. Moldoveanu, and E. A. Shudayfat, "Mixed reality to support new learning paradigms," in *Proc. 18th Int. Conf. Syst. Theory, Control Comput. (ICSTCC)*, Oct. 2014, pp. 692–697.

[46] M. A. González, B. S. N. Santos, A. R. Vargas, J. Martín-Gutiérrez, and A. R. Orihuela, "Virtual worlds. Opportunities and challenges in the 21st century," *Proc. Comput. Sci.*, vol. 25, pp. 330–337, Jan. 2013.

[47] T. Amorim, L. Tapparo, N. Marranghello, A. C. R. Silva, and A. S. Pereira, "A multiple intelligences theory-based 3D virtual lab environment for digital systems teaching," *Proc. Comput. Sci.*, vol. 29, pp. 1413–1422, Jan. 2014.

[48] K. Yoon, S.-K. Kim, J. J. Han, S. Han, and M. Preda, *MPEG-V: Bridging the Virtual and Real World*. New York, NY, USA: Academic, 2015.

[49] A. Moldoveanu, A. Gradinaru, O.-M. Ferche, and L. Stefan, "The 3D UPB mixed reality campus: Challenges of mixing the real and the virtual," in *Proc. 18th Int. Conf. Syst. Theory, Control Comput. (ICSTCC)*, Oct. 2014, pp. 538–543.
[50] T. Kwanya, S. Christine, and U. Peter, *Library 3.0: Intelligent Libraries and Apomediation*. Amsterdam, The Netherlands: Elsevier, 2014.
[51] D. M. Barry, N. Ogawa, A. Dharmawansa, H. Kanematsu, Y. Fukumura, T. Shirai, K. Yajima, and T. Kobayashi, "Evaluation for students' learning manner using eye blinking system in metaverse," *Proc. Comput. Sci.*, vol. 60, pp. 1195–1204, Jan. 2015.
[52] S.-V. Rehm, L. Goel, and M. Crespi, "The metaverse as mediator between technology, trends, and the digital transformation of society and business," *J. Virtual Worlds Res.*, vol. 8, no. 2, pp. 1–8, Oct. 2015.
[53] J. C. Chen, "The crossroads of English language learners, task-based instruction, and 3D multi-user virtual learning in second life," *Comput. Educ.*, vol. 102, pp. 152–171, Nov. 2016.
[54] A. Zackery, P. Shariatpanahi, M. M. Zolfagharzadeh, and A. A. Pourezzat, "Toward a simulated replica of futures: Classification and possible trajectories of simulation in futures studies," *Futures*, vol. 81, pp. 40–53, Aug. 2016.
[55] H. Kanematsu, N. Ogawa, A. Shimizu, T. Shirai, M. Kawaguchi, T. Kobayashi, K. T. Nakahira, and D. M. Barry, "Skype discussion for PBL between two laboratories and students biological/psychological responses," *Proc. Comput. Sci.*, vol. 112, pp. 1730–1736, Jan. 2017.
[56] K. J. L. Nevelsteen, "Virtual world, defined from a technological perspective and applied to video games, mixed reality, and the metaverse," *Comput. Animation Virtual Worlds*, vol. 29, no. 1, p. e1752, Jan. 2018.
[57] J. Huggett, "Virtually real or really virtual: Towards a heritage metaverse," *Stud. Digit. Heritage*, vol. 4, no. 1, pp. 1–15, Jun. 2020.
[58] P. De Decker and S. Peterson, "Beyond virtual or physical environments: Building a research metaverse a white paper for NDRIO's Canadian digital research needs assessment," Digit. Res. Alliance Canada, Toronto, ON, Canada, Tech. Rep., 2020.
[59] A. Siyaev and G.-S. Jo, "Towards aircraft maintenance metaverse using speech interactions with virtual objects in mixed reality," *Sensors*, vol. 21, no. 6, p. 2066, Mar. 2021.
[60] M. Dowling, "Fertile LAND: Pricing non-fungible tokens," *Finance Res. Lett.*, Apr. 2021, Art. no. 102096.
[61] *Metaverse Wiki*. Accessed: Nov. 11, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Metaverse
[62] N. Stephenson, *Snow Crash*. New York, NY, USA: Bantam Books, 1992.
[63] S. G. Lee, S. Trimi, W. K. Byun, and M. Kang, "Innovation and imitation effects in Metaverse service adoption," *Service Bus.*, vol. 5, no. 2, pp. 155–172, 2011.
[64] M. Grimshaw, *The Oxford Handbook of Virtuality*. New York, NY, USA: Oxford Univ. Press, 2014, p. 702.
[65] *Digital Twin Wiki*. Accessed: Nov. 11, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Digital_twin
[66] K. Oddone, "Even better than the real thing?" Virtual Augmented Reality School Library, SCIS Connections, Educ. Service Australia, Melbourne, VIC, Australia, Tech. Rep., 2019, pp. 1–15, no. 110.
[67] S. Townsdin and W. Whitmer, "Implementing augmented reality in academic libraries," *Public Services Quart.*, vol. 13, pp. 190–199, Jul. 2017, doi: 10.1080/15228959.2017.1338541.
[68] M. Alcañiz, E. Bigné, and J. Guixeres, "Virtual reality in marketing: A framework, review, and research agenda," *Frontiers Psychol.*, vol. 10, p. 1530, Jul. 2019, doi: 10.3389/fpsyg.2019.01530/.
[69] M. Alcañiz, E. Bigné, and J. Guixeres, "Virtual reality in marketing: A framework, review, and research agenda," *Frontiers Psychol.*, vol. 10, p. 1530, Jul. 2019.
[70] L. Birnie, T. Abhayapala, V. Tourbabin, and P. Samarasinghe, "Mixed source sound field translation for virtual binaural application with perceptual validation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1188–1203, 2021.
[71] J. W. Ruffner, J. E. Fulbrook, and M. Foglia, "Near-to-eye display concepts for air traffic controllers," *Proc. SPIE*, vol. 5442, pp. 120–131, Sep. 2004.
[72] Z. Li, J. Chan, J. Walton, H. Benko, D. Wigdor, and M. Glueck, "Armstrong: An empirical examination of pointing at non-dominant arm-anchored UIs in virtual reality," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–14.
[73] E. Bouzbib, G. Bailly, S. Haliyo, and P. Frey, "'Can i touch this?': Survey of virtual reality interactions via haptic solutions," 2021, *arXiv:2101.11278*.
[74] C. R. Foy, J. J. Dudley, A. Gupta, H. Benko, and P. O. Kristensson, "Understanding, detecting and mitigating the effects of coactivations in ten-finger mid-air typing in virtual reality," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–11.
[75] Z. Ren, I. Misra, A. G. Schwing, and R. Girdhar, "3D spatial recognition without spatially labeled 3D," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13204–13213.
[76] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang, "Factorizable Net: An efficient subgraph-based framework for scene graph generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 335–351.
[77] M. Chen, A. Thierry, and D. Ludovic, "Unsupervised object segmentation by redrawing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–18.
[78] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4693–4702.
[79] M. Fazel-Zarandi, S. Biswas, R. Summers, A. Elmalt, A. McCraw, M. McPhilips, and J. Peach, "Towards personalized dialog policies for conversational skill discovery," in *Proc. 33rd Conf. Neural Inf. Process. Syst. (NIPS)*, 2019.
[80] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5831–5840.
[81] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt, "XNect: Real-time multi-person 3D motion capture with a single RGB camera," *ACM Trans. Graph.*, vol. 39, no. 4, p. 82, Jul. 2020.
[82] M. Wang, X.-Q. Lyu, Y.-J. Li, and F.-L. Zhang, "VR content creation and exploration with deep learning: A survey," *Comput. Vis. Media*, vol. 6, no. 1, pp. 3–28, Mar. 2020.
[83] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017.
[84] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 290–298.
[85] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 6, Jul. 2017, p. 2.
[86] C. C. Park, B. Kim, and G. Kim, "Attend to you: Personalized image captioning with context sequence memory networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 895–903.
[87] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.
[88] L. Yang, K. Tang, J. Yang, and L.-J. Li, "Dense captioning with joint inference and visual context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR))*, Jul. 2017, pp. 1978–1987.
[89] R. Vedantam, S. Bengio, K. Murphy, D. Parikh, and G. Chechik, "Context-aware captions from context-agnostic supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 251–260.
[90] J. Fu, T. Mei, K. Yang, H. Lu, and Y. Rui, "Tagging personal photos with transfer deep learning," in *Proc. 24th Int. Conf. World Wide Web, Int. World Wide Web Conf. Steering Committee*, May 2015, pp. 344–354.
[91] P. Rastogi, A. Poliak, V. Lyzinski, and B. Van Durme, "Neural variational entity set expansion for automatically populated knowledge graphs," *Inf. Retr. J.*, vol. 22, nos. 3–4, pp. 232–255, Aug. 2019.
[92] Y. Fu, Y. Feng, and J. P. Cunningham, "Paraphrase generation with latent bag of words," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13623–13634.
[93] D. Zeng, H. Zhang, L. Xiang, J. Wang, and G. Ji, "User-oriented paraphrase generation with keywords controlled network," *IEEE Access*, vol. 7, pp. 80542–80551, 2019.
[94] Y. Ma, K. L. Nguyen, F. Z. Xing, and E. Cambria, "A survey on empathetic dialogue systems," *Inf. Fusion*, vol. 64, pp. 50–70, Dec. 2020.
[95] Y. Zheng, G. Chen, M. Huang, S. Liu, and X. Zhu, "Persona-aware dialogue generation with enriched profile," in *Proc. 33rd Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, 2019, pp. 1–10.
[96] N. Lubis, S. Sakti, K. Yoshino, and S. Nakamura, "Dialogue model and response generation for emotion improvement elicitation," in *Proc. 33rd Conf. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 1–11.

[97] K. Chandu, S. Prabhumoye, R. Salakhutdinov, and A. W. Black, "'My way of telling a story': Persona based grounded story generation," in *Proc. 2nd Workshop Storytelling*, 2019, pp. 11–21.
[98] *CTRL—A Conditional Transformer Language Model for Controllable Generation GitHub*. Accessed: Nov. 11, 2021. [Online]. Available: https://github.com/salesforce/ctrl
[99] C. Moon, P. Jones, and N. F. Samatova, "Learning entity type embeddings for knowledge graph completion," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 2181–2187.
[100] L. Poddar, W. Hsu, and M. L. Lee, "Author-aware aspect topic sentiment model to retrieve supporting opinions from reviews," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 472–481.
[101] M. Zhang, Y. Zhang, and G. Fu, "End-to-end neural relation extraction with global optimization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1730–1740.
[102] F. Liu, B. Liu, C. Sun, M. Liu, and X. Wang, "Multimodal learning-based approaches for link prediction in social networks," in *Natural Language Processing and Chinese Computing*. Cham, Switzerland: Springer, 2015, pp. 123–133.
[103] Q. Ning, Z. Feng, and D. Roth, "A structured learning approach to temporal relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1027–1037.
[104] Q. Huang, Z. Gan, A. Celikyilmaz, D. Wu, J. Wang, and X. He, "Hierarchically structured reinforcement learning for topically coherent visual story generation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8465–8472.
[105] F. Liu, B. Liu, C. Sun, M. Liu, and X. Wang, "Multimodal deep belief network based link prediction and user comment generation," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, Nov. 2015, pp. 20–28.
[106] C. Ciliberto, A. Rudi, L. Rosasco, and M. Pontil, "Consistent multitask learning with nonlinear output relations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1986–1996.
[107] W. Zhuo, M. Salzmann, X. He, and M. Liu, "Indoor scene parsing with instance segmentation, semantic labeling and support relationship inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5429–5437.
[108] P. K. Choubey and R. Huang, "A sequential model for classifying temporal relations between intra-sentence events," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1–7.
[109] S. Woo, D. Kim, D. Cho, and I. S. Kweon, "LinkNet: Relational embedding for scene graph," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1–11.
[110] L. Xiao, X. Hu, Y. Chen, Y. Xue, D. Gu, B. Chen, and T. Zhang, "Targeted sentiment classification based on attentional encoding and graph convolutional networks," *Appl. Sci.*, vol. 10, no. 3, p. 957, Feb. 2020.
[111] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1437–1445.
[112] E. Zuo, H. Zhao, B. Chen, and Q. Chen, "Context-specific heterogeneous graph convolutional network for implicit sentiment analysis," *IEEE Access*, vol. 8, pp. 37967–37975, 2020.
[113] M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath, "CausalGAN: Learning causal implicit generative models with adversarial training," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–37.
[114] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 299–307.
[115] X. Geng, H. Zhang, Z. Song, Y. Yang, H. Luan, and T.-S. Chua, "One of a kind: User profiling by social curation," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 567–576.
[116] L. Gui, J. Hu, Y. He, R. Xu, L. Qin, and J. Du, "A question answering approach for emotion cause extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1–10.
[117] F. Yang, Z. Yang, and W. W. Cohen, "Differentiable learning of logical rules for knowledge base reasoning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2319–2328.
[118] H. Lin, L. Sun, and X. Han, "Reasoning with heterogeneous knowledge for commonsense machine comprehension," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2032–2043.
[119] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh, "Cluster-GCN: An efficient algorithm for training deep and large graph convolutional networks," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 257–266.
[120] C. Sun, Y. Gong, Y. Wu, M. Gong, D. Jiang, M. Lan, S. Sun, and N. Duan, "Joint type inference on entities and relations via graph convolutional networks," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1361–1370.
[121] P. Tambwekar, M. Dhuliawala, L. J. Martin, A. Mehta, B. Harrison, and M. O. Riedl, "Controllable neural story plot generation via reward shaping," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1–7.
[122] M. Isonuma, T. Fujino, J. Mori, Y. Matsuo, and I. Sakata, "Extractive summarization using multi-task learning with document classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2101–2110.
[123] H. Yu and M. O. Riedl, "A sequential recommendation approach for interactive personalized story generation," in *Proc. 11th Int. Conf. Auto. Agents Multiagent Syst.*, vol. 1, Jun. 2012, pp. 71–78.
[124] C. Gulcehre, F. Dutil, A. Trischler, and Y. Bengio, "Plan, attend, generate: Planning for sequence-to-sequence models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5474–5483.
[125] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–37.
[126] C. J. Nan, K. M. Kim, and B. T. Zhang, "Social network analysis of TV drama characters via deep concept hierarchies," in *Proc. Adv. Social Netw. Anal. Mining (ASONAM), IEEE/ACM Int. Conf.*, Aug. 2015, pp. 831–836.
[127] A. Newell and J. Deng, "Pixels to graphs by associative embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2168–2177.
[128] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori, "Learning structured inference neural networks with label relations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2960–2968.
[129] C. Xiang, T. Jiang, B. Chang, and Z. Sui, "ERSOM: A structural ontology matching approach using automatically learned entity representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2419–2429.
[130] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 452–461.
[131] V. Niculae and M. Blondel, "A regularized framework for sparse and structured neural attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3338–3348.
[132] L. Baraldi, C. Grana, and R. Cucchiara, "Recognizing and presenting the storytelling video structure with deep multimodal networks," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 955–968, May 2017.
[133] M. Bolaños, M. Dimiccoli, and P. Radeva, "Toward storytelling from visual lifelogging: An overview," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 1, pp. 77–90, Feb. 2017.
[134] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, and J. Gao, "StoryGAN: A sequential conditional GAN for story visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6329–6338.
[135] H. Zhang, Z. Kyaw, S. F. Chang, and T. S. Chua, "Visual translation embedding network for visual relation detection," in *Proc. CVPR*, vol. 1, no. 2, Jul. 2017, p. 5.
[136] L. Yu, M. Bansal, and T. Berg, "Hierarchically-attentive RNN for album summarization and storytelling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1–7.
[137] Q. Sun, S. Lee, and D. Batra, "Bidirectional beam search: Forward-backward inference in neural sequence models for fill-in-the-blank image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6961–6969.
[138] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling relationships in referential expressions with compositional modular networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4418–4427.
[139] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang, "Video captioning via hierarchical reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4213–4222.
[140] J. Tang, "Show, reward, and tell: Adversarial visual story generation," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 2s, pp. 1–20, 2019.
[141] J. Novikova, O. Dušek, A. C. Curry, and V. Rieser, "Why we need new evaluation metrics for NLG," *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, Sep. 2017, pp. 2231–2242.
[142] L. Huang, K. Zhao, and M. Ma, "When to finish? Optimal beam search for neural text generation (modulo beam size)," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1–6.
[143] J. Eisenberg and M. Finlayson, "A simpler and more generalizable story detector using verb and character features," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2708–2715.

[144] S. Swayamdipta, A. P. Parikh, and T. Kwiatkowski, "Multi-mention learning for reading comprehension with neural cascades," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–12.
[145] G. A. B. Barros, M. C. Green, A. Liapis, and J. Togelius, "Who killed Albert Einstein? From open data to murder mystery games," *IEEE Trans. Games*, vol. 11, no. 1, pp. 79–89, Mar. 2019.
[146] L. Bounegru, T. Venturini, J. Gray, and M. Jacomy, "Narrating networks: Exploring the affordances of networks as storytelling devices in journalism," *Digit. J.*, vol. 5, no. 6, pp. 699–730, 2017.
[147] W. Y. Wang, Y. Mehdad, D. R. Radev, and A. Stent, "A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 58–68.
[148] L. Shi and R. Setchi, "User-oriented ontology-based clustering of stored memories," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9730–9742, Aug. 2012.
[149] Y.-G. Cheong, Y.-J. Kim, W.-H. Min, E.-S. Shim, and J.-Y. Kim, "Prism: A framework for authoring interactive narratives," in *Proc. Joint Int. Conf. Interact. Digit. Storytelling*. Berlin, Germany: Springer, 2008, pp. 297–308.
[150] I. Subašić and B. Bettina, "Experience STORIES: A visual news search and summarization system," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Berlin, Germany: Springer, 2010, pp. 619–623.
[151] M. Krstajić, M. Najm-Araghi, F. Mansmann, and D. A. Keim, "Story tracker: Incremental visual text analytics of news story development," *Inf. Visualizat.*, vol. 12, nos. 3–4, pp. 308–323, Jul. 2013.
[152] A. Miller, W. Feng, D. Batra, A. Bordes, A. Fisch, J. Lu, D. Parikh, and J. Weston, "ParlAI: A dialog research software platform," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2017, pp. 1–7.
[153] A. Das, S. Kottur, J. M. F. Moura, S. Lee, and D. Batra, "Learning cooperative visual dialog agents with deep reinforcement learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2951–2960.
[154] J. A. Mendez, A. Geramifard, M. Ghavamzadeh, and B. Liu, "Reinforcement learning of multi-domain dialog policies via action embeddings," in *Proc. 3rd Workshop Conversational AI Today's Practice Tomorrow's Potential (NIPS)*, 2019, pp. 1–11.
[155] J. Wu, G. Li, S. Liu, and L. Lin, "Tree-structured policy based progressive reinforcement learning for temporally language grounding in video," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, Apr. 2020, pp. 12386–12393.
[156] G. Gkioxari, R. Girshick, P. Dollar, and K. He, "Detecting and recognizing human-object interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8359–8367.
[157] K. Nguyen and H. Daumé III, "Help, anna! Visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 1–18.
[158] Y. Jiang, S. Gu, K. Murphy, and C. Finn, "Language as an abstraction for hierarchical deep reinforcement learning," in *Proc. Advances Neural Inf. Process. Syst.*, 2019, pp. 1–25.
[159] S. W. Lee, T. Gao, S. Yang, J. Yoo, and J. W. Ha, "Large-scale answerer in questioner's mind for visual dialog question generation," in *Proc. ICLR*, 2019, pp. 1–16.
[160] P. Ammanabrolu and M. Hausknecht, "Graph constrained reinforcement learning for natural language action spaces," in *Proc. ICRL*, 2020, pp. 1–24.
[161] T. Domhan and F. Hieber, "Using target-side monolingual data for neural machine translation through multi-task learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1500–1505.
[162] D. Camacho, Á. Panizo-LLedot, G. Bello-Orgaz, A. Gonzalez-Pardo, and E. Cambria, "The four dimensions of social network analysis: An overview of research methods, applications, and software tools," *Inf. Fusion*, vol. 63, pp. 88–120, Nov. 2020.
[163] J. Kruk, J. Lubin, K. Sikka, X. Lin, D. Jurafsky, and A. Divakaran, "Integrating text and image: Determining multimodal document intent in Instagram posts," *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 4622–4632.
[164] S. Palaskar, J. Libovický, S. Gella, and F. Metze, "Multimodal abstractive summarization for How2 videos," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1–10.
[165] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, Apr. 2020, pp. 13041–13049.
[166] M. Hessel, H. Soyer, L. Espeholt, W. Czarnecki, S. Schmitt, and H. van Hasselt, "Multi-task deep reinforcement learning with popart," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3796–3803.
[167] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," in *Proc. Interspeech*, Sep. 2019, pp. 1–5.
[168] Y. Qian, R. Ubale, V. Ramanaryanan, P. Lange, D. Suendermann-Oeft, K. Evanini, and E. Tsuprun, "Exploring ASR-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system," in *Proc. IEEE Automat. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 569–576.
[169] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1–10.
[170] M. Poesio, S. Roland, and V. Yannick, *Anaphora Resolution*. Cham, Switzerland: Springer, 2016.
[171] S. Shekhar, U. Kumar, and U. Sharma, "To reduce the multidimensionality of feature set for anaphora resolution algorithm," in *Ambient Communications and Computer Systems*. Singapore: Springer, 2018, pp. 437–446.
[172] Y. Niu, H. Zhang, M. Zhang, J. Zhang, Z. Lu, and J.-R. Wen, "Recursive visual attention in visual dialog," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6679–6688.
[173] R. Sukthanker, S. Poria, E. Cambria, and R. Thirunavukarasu, "Anaphora and coreference resolution: A review," *Inf. Fusion*, vol. 59, pp. 139–162, Jul. 2020.
[174] A. Rohrbach, M. Rohrbach, S. Tang, S. J. Oh, and B. Schiele, "Generating descriptions with grounded and co-referenced people," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4979–4989.
[175] B. Aktaş, T. Scheffler, and M. Stede, "Anaphora resolution for Twitter conversations: An exploratory study," in *Proc. 1st Workshop Comput. Models Reference, Anaphora Coreference*, 2018, pp. 1–10.
[176] S. Heinrich, M. Kerzel, E. Strahl, and S. Wermter, "Embodied multimodal interaction in language learning: The emil data collection," in *Proc. ICDL-EpiRob Workshop Active Vis., Attention, Learn. (ICDL-Epirob AVAL)(Tokyo)*, vol. 2, 2018, pp. 1–2.
[177] Y. Jiang, W. Li, M. S. Hossain, M. Chen, A. Alelaiwi, and M. Al-Hammadi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," *Inf. Fusion*, vol. 53, pp. 209–221, Jan. 2020.
[178] J. Zhang, K. Shih, A. Tao, B. Catanzaro, and A. Elgammal, "An interpretable model for scene graph generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1–6.
[179] P. P. Liang, A. Zadeh, and L.-P. Morency, "Multimodal local-global ranking fusion for emotion recognition," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 472–476.
[180] K. Ethayarajh, "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 1–11.
[181] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: A compact task-agnostic BERT for resource-limited devices," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1–13.
[182] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and S. Agarwal, "Language models are few-shot learners," 2020, *arXiv:2005.14165*.
[183] H. Tan and M. Bansal, "Vokenization: Improving language understanding with contextualized, visual-grounded supervision," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 1–15.
[184] *T5: Text-to-Text Transfer Transformer GitHub*. Accessed: Nov. 11, 2021. [Online]. Available: https:github.com/google-research/text-to-text-transfer-transformer
[185] H. Le, D. Sahoo, N. Chen, and S. Hoi, "Multimodal transformer networks for end-to-end video-grounded dialogue systems," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5612–5623.
[186] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–11.
[187] Z. Zheng, W. Wang, S. Qi, and S.-C. Zhu, "Reasoning visual dialogs with structural and partial observations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6669–6678.

[188] A. S. Gordon, C. A. Bejan, and K. Sagae, "Commonsense causal reasoning using millions of personal stories," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2011, pp. 1–6.
[189] D. Wang, M. Jamnik, and P. Lio, "Abstract diagrammatic reasoning with multiplex graph networks," in *Proc. ICRL*, 2020, pp. 1–20.
[190] A. Asai, K. Hashimoto, H. Hajishirzi, R. Socher, and C. Xiong, "Learning to retrieve reasoning paths over wikipedia graph for question answering," in *Proc. ICLR*, 2020, pp. 1–22.
[191] W. Xiong, T. Hoang, and W. Y. Wang, "DeepPath: A reinforcement learning method for knowledge graph reasoning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1–10.
[192] H. Zhu, Y. Lin, Z. Liu, J. Fu, T.-S. Chua, and M. Sun, "Graph neural networks with generated parameters for relation extraction," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1331–1339.
[193] H. Zeng, H. Zhou, A. Srivastava, R. Kannan, and V. Prasanna, "Graphsaint: Graph sampling based inductive learning method," in *Proc. ICRL*, 2020, pp. 1–19.
[194] T. Bansal, D.-C. Juan, S. Ravi, and A. McCallum, "A2N: Attending to neighbors for knowledge graph inference," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4387–4392.
[195] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, "Diversity is all you need: Learning skills without a reward function," in *Proc. ICLR*, 2019, pp. 1–22.
[196] M. Laskin, A. Srinivas, and P. Abbeel, "CURL: Contrastive unsupervised representations for reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1–20.
[197] X. Puig, T. Shu, S. Li, Z. Wang, Y.-H. Liao, J. B. Tenenbaum, S. Fidler, and A. Torralba, "Watch- and-help: A challenge for social perception and human-AI collaboration," in *Proc. ICLR*, 2021, pp. 1–23.
[198] J. B. Hamrick, K. R. Allen, V. Bapst, T. Zhu, K. R. McKee, J. B. Tenenbaum, and P. W. Battaglia, "Relational inductive bias for physical construction in humans and machines," in *Proc. Annu. Meeting Cogn. Sci. Soc. (CogSci)*, 2018, pp. 1–7.
[199] J. B. Hamrick, "Analogues of mental simulation and imagination in deep learning," *Current Opinion Behav. Sci.*, vol. 29, pp. 8–16, Oct. 2019.
[200] C. Davis, L. Bulat, A. Vero, and E. Shutova, "Modelling visual properties and visual context in multimodal semantics," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1–6.
[201] W. Dabney, Z. Kurth-Nelson, N. Uchida, C. K. Starkweather, D. Hassabis, R. Munos, and M. Botvinick, "A distributional code for value in dopamine-based reinforcement learning," *Nature*, vol. 577, pp. 1–5, Jan. 2020.
[202] M. Moscovitch, R. Cabeza, G. Winocur, and L. Nadel, "Episodic memory and beyond: The hippocampus and neocortex in transformation," *Annu. Rev. Psychol.*, vol. 67, pp. 105–134, Jan. 2016.
[203] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6467–6476.
[204] P. Y. Oudeyer, J. Gottlieb, and M. Lopes, "Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies," *Prog. Brain Res.*, vol. 229, pp. 257–284, Jan. 2016.
[205] N. C. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick, "Machine theory of mind," in *Proc. 35th Int. Conf. Mach. Learn.*, Stockholm, Sweden, vol. 80, 2018, pp. 4218–4227.
[206] D. Melhart, G. N. Yannakakis, and A. Liapis, "I feel i feel you: A theory of mind experiment in games," *KI Künstliche Intelligenz*, vol. 34, no. 1, pp. 45–55, Mar. 2020.
[207] A. Raj, J. Tanke, J. Hays, M. Vo, C. Stoll, and C. Lassner, "ANR: Articulated neural rendering for virtual avatars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3722–3731.
[208] L. Liu, T. Zhou, G. Long, J. Jiang, and C. Zhang, "Learning to propagate for graph meta-learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–12.
[209] R. Vuorio, S. H. Sun, H. Hu, and J. J. Lim, "Multimodal model-agnostic meta-learning via task-aware modulation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–22.
[210] S. Racanière, T. Weber, D. P. Reichert, L. Buesing, A. Guez, D. Rezende, A. P. Badia, O. Vinyals, N. Heess, Y. Li, and R. Pascanu, "Imagination-augmented agents for deep reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5694–5705.
[211] R. Skarbez, M. Smith, and M. C. Whitton, "Revisiting milgram and Kishino's reality-virtuality continuum," *Frontiers Virtual Reality*, vol. 2, p. 27, Mar. 2021.
[212] P. Maharg and M. Owen, "Simulations, learning and the metaverse: Changing cultures in legal education," *J. Inf., Law, Technol.*, vol. 1, pp. 1–28, Jan. 2007.
[213] J. Shi, T. Honjo, K. Zhang, and K. Furuya, "Using virtual reality to assess landscape: A comparative study between on-site survey and virtual reality of aesthetic preference and landscape cognition," *Sustainability*, vol. 12, no. 7, p. 2875, Apr. 2020.
[214] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "IQA: Visual question answering in interactive environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4089–4098.
[215] Y. Qiu, A. Pal, and H. I. Christensen, "Target driven visual navigation exploiting object relationships," Tech. Rep., Mar. 2020. Accessed: Nov. 1, 2021. [Online]. Available: https://arxiv.org/pdf/2003.06749v1.pdf
[216] J. Li, S. Tang, F. Wu, and Y. Zhuang, "Walking with MIND: Mental imagery eNhanceD embodied QA," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1211–1219.
[217] R. Tamari, C. Shani, T. Hope, M. R. L. Petruck, O. Abend, and D. Shahaf, "Language (Re)modelling: Towards embodied language understanding," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1–14.
[218] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch, "Emergent tool use from multi-agent autocurricula," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–28.
[219] N. Ilinykh, S. Zarrieß, and D. Schlangen, "Meetup! A corpus of joint activity dialogues in a visual environment," in *Proc. 23rd Workshop Semantics Pragmatics Dialogue (Semdial/LondonLogue)*, London, U.K., Sep. 2019, pp. 1–10.
[220] B. Chen, S. Song, H. Lipson, and C. Vondrick, *Visual Hide and Seek* (Artificial Life Conference Proceedings). Cambridge, MA, USA: MIT Press, 2020.
[221] I.-C. Stanica, F. Moldoveanu, G.-P. Portelli, M.-I. Dascalu, A. Moldoveanu, and M. G. Ristea, "Flexible virtual reality system for neurorehabilitation and quality of life improvement," *Sensors*, vol. 20, no. 21, p. 6045, Oct. 2020.
[222] S. Papagiannidis, M. Bourlakis, and F. Li, "Making real money in virtual worlds: MMORPGs and emerging business opportunities, challenges and ethical implications in Metaverses," *Technol. Forecasting Social Change*, vol. 75, no. 5, pp. 610–622, 2008.
[223] J. Smart, J. Cascio, J. Paffendorf, C. Bridges, J. Hummel, J. Hursthouse, and R. Moss, "A cross-industry public foresight project," in *Proc. Metaverse Roadmap Pathways 3DWeb*, 2007, pp. 1–28.
[224] Y. Tang, "Help first-year college students to learn their library through an augmented reality game," *J. Academic Librarianship*, vol. 47, no. 1, Jan. 2021, Art. no. 102294.
[225] M. Noghabaei, A. Heydarian, V. Balali, and K. Han, "A survey study to understand industry vision for virtual and augmented reality applications in design and construction," 2020, *arXiv:2005.02795*.
[226] H. Kanematsu, T. Kobayashi, D. M. Barry, Y. Fukumura, A. Dharmawansa, and N. Ogawa, "Virtual STEM class for nuclear safety education in metaverse," *Proc. Comput. Sci.*, vol. 35, pp. 1255–1261, Jan. 2014.
[227] B. Sung, E. Mergelsberg, M. Teah, B. D'Silva, and I. Phau, "The effectiveness of a marketing virtual reality learning simulation: A quantitative survey with psychophysiological measures," *Brit. J. Educ. Technol.*, vol. 52, no. 1, pp. 196–213, Jan. 2021.
[228] T. Templeton, "Getting real: Learning with (and about) augmented reality," *Scan, J. Educators*, vol. 39, no. 10, pp. 6–15, 2020.
[229] D. M. Barry, H. Kanematsu, Y. Fukumura, N. Ogawa, A. Okuda, R. Taguchi, and H. Nagai, "International comparison for problem based learning in metaverse," in *Proc. ICEE ICEER*, vol. 6066, 2009, pp. 1–7.
[230] H. Kanematsu, Y. Fukumura, N. Ogawa, A. Okuda, R. Taguchi, and H. Nagai, "Practice and evaluation of problem based learning in metaverse," in *Proc. EdMediaC Innovate Learn. Assoc. Advancement Comput. Educ. (AACE)*, 2009, pp. 2862–2870.
[231] N. Khan, K. Muhammad, T. Hussain, M. Nasir, M. Munsif, A. S. Imran, and M. Sajjad, "An adaptive game-based learning strategy for children road safety education and practice in virtual space," *Sensors*, vol. 21, no. 11, p. 3661, May 2021.
[232] Afnan, K. Muhammad, N. Khan, M.-Y. Lee, A. Imran, and M. Sajjad, "School of the future: A comprehensive study on the effectiveness of augmented reality as a tool for primary school Children's education," *Appl. Sci.*, vol. 11, no. 11, p. 5277, Jun. 2021.
[233] Y. Li, T. Nagarajan, B. Xiong, and K. Grauman, "Ego-Exo: Transferring visual representations from third-person to first-person videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6943–6953.

[234] W. Xian, J.-B. Huang, J. Kopf, and C. Kim, "Space-time neural irradiance fields for free-viewpoint video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9421–9431.
[235] R. Bonatti, A. Bucker, S. Scherer, M. Mukadam, and J. Hodgins, "Batteries, camera, action! Learning a semantic control space for expressive robot cinematography," 2020, *arXiv:2011.10118*.
[236] H. Zhang, Y. Ye, T. Shiratori, and T. Komura, "ManipNet: Neural manipulation synthesis with a hand-object spatial representation," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–14, Aug. 2021.
[237] C. Rognon, T. Bunge, M. Gao, C. Connor, B. Stephens-Fripp, C. Brown, and A. Israr, "An online survey on the perception of mediated social touch interaction and device design," 2021, *arXiv:2104.00086*.
[238] E. D. Z. Chase, A. Israr, P. Preechayasomboon, S. Sykes, A. Gupta, and J. Hartcher-O'Brien, "Learning vibes: Communication bandwidth of a single wrist-worn vibrotactile actuator," in *Proc. IEEE World Haptics Conf. (WHC)*, Jul. 2021, pp. 421–426.
[239] B. Stephens-Fripp, A. Israr, and C. Rognon, "A multichannel pneumatic analog control system for haptic displays: Multichannel pneumatic analog control system (MPACS)," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–7.
[240] Y. Yuan, S.-E. Wei, T. Simon, K. Kitani, and J. Saragih, "SimPoE: Simulated character control for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7159–7169.
[241] N. Neverova, A. Sanakoyeu, P. Labatut, D. Novotny, and A. Vedaldi, "Discovering relationships between object categories via universal canonical maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 404–413.
[242] B. Lucas and B. Kozary, "Chromatic-luminance and true-texture analysis for image mining in the social sciences," Meta, CA, USA, Tech. Rep., Mar. 2021. Accessed: Nov. 1, 2021. [Online]. Available: https://research.facebook.com/publications/chromatic-luminance-and-true-texture-analysis-for-image-mining-in-the-social-sciences/
[243] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun, "ConViT: Improving vision transformers with soft convolutional inductive biases," 2021, *arXiv:2103.10697*.
[244] L. Porzi, S. R. Bulo, and P. Kontschieder, "Improving panoptic segmentation at all scales," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7302–7311.
[245] B. Cheng, R. Girshick, P. Dollar, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15334–15342.
[246] Y. Jin, A. Patney, and A. Bovik, "Evaluating foveated video quality using entropic differencing," 2021, *arXiv:2106.06817*.
[247] B. Xiong, H. Fan, K. Grauman, and C. Feichtenhofer, "Multiview pseudo-labeling for semi-supervised learning from video," 2021, *arXiv:2104.00682*.
[248] J. Huang, G. Pang, R. Kovvuri, M. Toh, K. J. Liang, P. Krishnan, X. Yin, and T. Hassner, "A multiplexed network for end-to-end, multilingual OCR," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4547–4557.
[249] A. Singh, G. Pang, M. Toh, J. Huang, W. Galuba, and T. Hassner, "TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8802–8812.
[250] S. Sodhani, A. Zhang, and J. Pineau, "Multi-task reinforcement learning with context-based representations," 2021, *arXiv:2102.06177*.
[251] P. Henzler, J. Reizenstein, P. Labatut, R. Shapovalov, T. Ritschel, A. Vedaldi, and D. Novotny, "Unsupervised learning of 3D object categories from videos in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4700–4709.
[252] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda, "Unbiased teacher for semi-supervised object detection," 2021, *arXiv:2102.09480*.
[253] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.
[254] Y. Tian, X. Chen, and S. Ganguli, "Understanding self-supervised learning dynamics without contrastive pairs," 2021, *arXiv:2102.06810*.
[255] J. Donley, V. Tourbabin, B. Rafaely, and R. Mehra, "Adaptive multichannel signal enhancement based on multi-source contribution estimation," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 276–280.
[256] P. Tzirakis, A. Kumar, and J. Donley, "Multi-channel speech enhancement using graph neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 3415–3419.
[257] H. Helmholz, J. Ahrens, D. L. Alon, S. V. A. Gari, and R. Mehra, "Evaluation of sensor self-noise in binaural rendering of spherical microphone array signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 161–165.
[258] S. E. Chazan, L. Wolf, E. Nachmani, and Y. Adi, "Single channel voice separation for unknown number of speakers under reverberant and noisy settings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 3730–3734.
[259] T. Shlomo and B. Rafaely, "Blind amplitude estimation of early room reflections using alternating least squares," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 476–480.
[260] T. Shlomo and B. Rafaely, "Blind localization of early room reflections using phase aligned spatial correlation," *IEEE Trans. Signal Process.*, vol. 69, pp. 1213–1225, 2021.
[261] Y. Zhou, H. Jiang, and V. K. Ithapu, "On the predictability of HRTFs from ear shapes using deep networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 441–445.
[262] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Binaural reproduction based on bilateral ambisonics and ear-aligned HRTFs," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 901–913, 2021.
[263] C. Bartlomiej, T. C. Sang-Ik, and M. Ravish, "Full range omnidirectional sound source for near-field head-related transfer-functions measurement," *J. Audio Eng. Soc.*, vol. 69, no. 5, pp. 323–339, May 2021.
[264] S. A. Gari, J. M. Arend, P. Calamia, and P. Robinson, "Optimizing the spatial decomposition method for binaural rendering," *J. Audio Eng. Soc.*, vol. 68, no. 12, pp. 959–976, Dec. 2020.
[265] S. Ge, V. Goswami, C. Lawrence Zitnick, and D. Parikh, "Creative sketch generation," 2020, *arXiv:2011.10039*.
[266] B. Roziere, N. C. Rakotonirina, V. Hosu, A. Rasoanaivo, H. Lin, C. Couprie, and O. Teytaud, "Tarsier: Evolving noise injection in super-resolution GANs," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7028–7035.
[267] C. Lassner and M. Zollhofer, "Pulsar: Efficient sphere-based neural rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1440–1449.
[268] C. Wu, Z. Xiu, Y. Shi, O. Kalinli, C. Fuegen, T. Koehler, and Q. He, "Transformer-based acoustic modeling for streaming speech synthesis," in *Proc. Interspeech*, 2021, pp. 146–150.
[269] A. Richard, D. Markovic, I. D. Gebru, S. Krenn, G. A. Butler, F. Torre, and Y. Sheikh, "Neural synthesis of binaural speech from mono audio," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–23.
[270] J. D. Won, "Control strategies for physically simulated characters performing two-player competitive sports," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–11, 2021.
[271] Y. Ye, S. Tulsiani, and A. Gupta, "Shelf-supervised mesh prediction in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8843–8852.
[272] W. Yuan, Z. Lv, T. Schmidt, and S. Lovegrove, "STaR: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13144–13152.
[273] E. Ng, S. Ginosar, T. Darrell, and H. Joo, "Body2Hands: Learning to infer 3D hands from conversational gesture body dynamics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11865–11874.
[274] M. Eisenberger, D. Novotny, G. Kerchenbaum, P. Labatut, N. Neverova, D. Cremers, and A. Vedaldi, "NeuroMorph: Unsupervised shape interpolation and correspondence in one go," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7473–7483.
[275] B.-C. Chen, Z. Wu, L. S. Davis, and S.-N. Lim, "Efficient object embedding for spliced image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14965–14975.
[276] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, Z. Ma, T. Thrush, S. Riedel, Z. Waseem, P. Stenetorp, R. Jia, M. Bansal, C. Potts, and A. Williams, "Dynabench: Rethinking benchmarking in NLP," 2021, *arXiv:2104.14337*.
[277] T. Blevins, M. Joshi, and L. Zettlemoyer, "FEWS: Large-scale, low-shot word sense disambiguation with the dictionary," 2021, *arXiv:2102.07983*.
[278] N. De Cao, G. Izacard, S. Riedel, and F. Petroni, "Autoregressive entity retrieval," 2020, *arXiv:2010.00904*.

[279] O. Gafni, O. Ashual, and L. Wolf, "Single-shot freestyle dance reenactment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 882–891.
[280] C. Gong, D. Wang, M. Li, V. Chandra, and Q. Liu, "KeepAugment: A simple information-preserving data augmentation approach," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 1055–1064.
[281] B. Li, F. Wu, S.-N. Lim, S. Belongie, and K. Q. Weinberger, "On feature normalization and data augmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12383–12392.
[282] K. J Liang, W. Hao, D. Shen, Y. Zhou, W. Chen, C. Chen, and L. Carin, "MixKD: Towards efficient distillation of large-scale language models," 2020, *arXiv:2011.00593*.
[283] M. Jia, Z. Wu, A. Reiter, C. Cardie, S. Belongie, and S.-N. Lim, "Intentonomy: A dataset and study towards human intent understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12986–12996.
[284] Q. Huang, H. He, A. Singh, S.-N. Lim, and A. R. Benson, "Combining label propagation and simple models out-performs graph neural networks," 2020, *arXiv:2010.13993*.
[285] J. Ahlgren, K. Bojarczuk, S. Drossopoulou, I. Dvortsova, J. George, N. Gucevska, M. Harman, M. Lomeli, S. M. Lucas, E. Meijer, and S. Omohundro, "Facebook's cyber–cyber and cyber–physical digital twins," in *Evaluation and Assessment in Software Engineering*. Trondheim, Norway: Norwegian Univ. Science and Technology (NTNU), 2021, pp. 1–9.
[286] *BlenderBot 2.0 GitHub*. Accessed: Nov. 11, 2021. [Online]. Available: https://github.com/facebookresearch/ParlAI/tree/master/projects/blenderbot2
[287] G. Dagan, D. Hupkes, and E. Bruni, "Co-evolution of language and agents in referential games," 2020, *arXiv:2001.03361*.
[288] D. Le, G. Keren, J. Chan, J. Mahadeokar, C. Fuegen, and M. L. Seltzer, "Deep shallow fusion for RNN-T personalization," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 251–257.
[289] L. Weber, J. Jumelet, E. Bruni, and D. Hupkes, "Language modelling as a multi-task problem," 2021, *arXiv:2101.11287*.
[290] P. Lewis, B. Oğuz, R. Rinott, S. Riedel, and H. Schwenk, "MLQA: Evaluating cross-lingual extractive question answering," 2019, *arXiv:1910.07475*.
[291] W. Xiong, X. Lorraine Li, S. Iyer, J. Du, P. Lewis, W. Y. Wang, Y. Mehdad, W.-T. Yih, S. Riedel, D. Kiela, and B. Oğuz, "Answering complex open-domain questions with multi-hop dense retrieval," 2020, *arXiv:2009.12756*.
[292] S. Min, J. Boyd-Graber, C. Alberti, D. Chen, E. Choi, M. Collins, K. Guu, H. Hajishirzi, K. Lee, J. Palomaki, and C. Raffel, "NeurIPS 2020 EfficientQA competition: Systems, analyses and lessons learned," 2021, *arXiv:2101.00133*.
[293] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, and F. Guzmán, "WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia," 2019, *arXiv:1907.05791*.
[294] Y. Tang, J. Pino, X. Li, C. Wang, and D. Genzel, "Improving speech translation by understanding and learning from the auxiliary text translation task," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 4252–4261.
[295] Y.-L. Tuan, A. El-Kishky, A. Renduchintala, V. Chaudhary, F. Guzmán, and L. Specia, "Quality estimation without human-labeled data," 2021, *arXiv:2102.04020*.
[296] M. Patrick, P.-Y. Huang, Y. Asano, F. Metze, A. Hauptmann, J. Henriques, and A. Vedaldi, "Support-set bottlenecks for video-text representation learning," 2020, *arXiv:2010.02824*.
[297] Z. Meng, L. Yu, N. Zhang, T. Berg, B. Damavandi, V. Singh, and A. Bearman, "Connecting what to say with where to look by modeling human attention traces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12679–12688.
[298] P. Morgado, I. Misra, and N. Vasconcelos, "Robust audio-visual instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12934–12945.
[299] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12475–12486.
[300] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. S. Chaplot, O. Maksymets, and A. Gokaslan, "Habitat 2.0: Training home assistants to rearrange their habitat," 2021, *arXiv:2106.14405*.
[301] C. Chen, S. Majumder, Z. Al-Halah, R. Gao, S. Kumar Ramakrishnan, and K. Grauman, "Learning to set waypoints for audio-visual navigation," 2020, *arXiv:2008.09622*.
[302] C. Chen, Z. Al-Halah, and K. Grauman, "Semantic audio-visual navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15516–15525.
[303] J. Ye, D. Batra, A. Das, and E. Wijmans, "Auxiliary tasks and exploration enable ObjectNav," 2021, *arXiv:2104.04112*.
[304] T. Bagautdinov, C. Wu, T. Simon, F. Prada, T. Shiratori, S.-E. Wei, W. Xu, Y. Sheikh, and J. Saragih, "Driving-signal aware full-body avatars," 2021, *arXiv:2105.10441*.
[305] S. Lombardi, T. Simon, G. Schwartz, M. Zollhoefer, Y. Sheikh, and J. Saragih, "Mixture of volumetric primitives for efficient neural rendering," 2021, *arXiv:2103.01954*.
[306] T. Sun, G. Nam, C. Aliaga, C. Hery, and R. Ramamoorthi, "Human hair inverse rendering using multi-view photometric data," in *Proc. Eurograph. Symp. Rendering (EGSR)*, 2021, pp. 179–190.
[307] D. Xiang, F. Prada, T. Bagautdinov, W. Xu, Y. Dong, H. Wen, J. Hodgins, and C. Wu, "Modeling clothing as a separate layer for an animatable human avatar," 2021, *arXiv:2106.14879*.
[308] B. Chaudhuri, N. Sarafianos, L. Shapiro, and T. Tung, "Semi-supervised synthesis of high-resolution editable textures for 3D humans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7991–8000.
[309] S. Bi, S. Lombardi, S. Saito, T. Simon, S. E. Wei, K. McPhail, R. Ramamoorthi, Y. Sheikh, and J. Saragih, "Deep relightable appearance models for animatable faces," *ACM Trans. Graph.*, vol. 30, pp. 1–15, Jul. 2021.
[310] S. Ma, T. Simon, J. Saragih, D. Wang, Y. Li, F. D. La Torre, and Y. Sheikh, "Pixel codec avatars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 64–73.
[311] A. Shukla, S. Petridis, and M. Pantic, "Does visual self-supervision improve learning of speech representations for emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Feb. 26, 2021, doi: 10.1109/TAFFC.2021.3062406.
[312] M. Jiang, E. Grefenstette, and T. Rocktäschel, "Prioritized level replay," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4940–4950.
[313] N. Modhe, P. Chattopadhyay, M. Sharma, A. Das, D. Parikh, D. Batra, and R. Vedantam, "IR-VIC: Unsupervised discovery of sub-goals for transfer in RL," 2019, *arXiv:1907.10580*.
[314] B. Cui, Y. Chow, and M. Ghavamzadeh, "Control-aware representations for model-based reinforcement learning," 2020, *arXiv:2006.13408*.
[315] B. Amos, S. Stanton, D. Yarats, and A. G. Wilson, "On the model-based stochastic value gradient for continuous reinforcement learning," in *Learning for Dynamics and Control*. Cambridge, U.K.: Proceedings of Machine Learning Research, 2021.
[316] S. Sukhbaatar, D. Ju, S. Poff, S. Roller, A. Szlam, J. Weston, and A. Fan, "Not all memories are created equal: Learning to forget by expiring," 2021, *arXiv:2105.06548*.
[317] M. Abdelsalam, M. Faramarzi, S. Sodhani, and S. Chandar, "IIRC: Incremental implicitly-refined classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11038–11047.
[318] V. K. Verma, K. J. Liang, N. Mehta, P. Rai, and L. Carin, "Efficient feature transformations for discriminative and generative continual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13865–13875.
[319] N. Mehta, K. Liang, V. K. Verma, and L. Carin, "Continual learning using a Bayesian nonparametric dictionary of weight factors," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 100–108.
[320] S. Ebrahimi, S. Petryk, A. Gokul, W. Gan, J. E. Gonzalez, M. Rohrbach, and T. Darrell, "Remembering for the right reasons: Explanations reduce catastrophic forgetting," 2020, *arXiv:2010.01528*.
[321] S. Sodhani, A. Zhang, and J. Pineau, "Multi-task reinforcement learning with context-based representations," 2021, *arXiv:2102.06177*.
[322] C. Fu, H. Huang, X. Chen, Y. Tian, and J. Zhao, "Learn-to-share: A hardware-friendly transfer learning framework exploiting computation and parameter sharing," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 3469–3479.
[323] A. Zhang, S. Sodhani, K. Khetarpal, and J. Pineau, "Learning robust state abstractions for hidden-parameter block MDPs," 2020, *arXiv:2007.07206*.
[324] P. Dollar, M. Singh, and R. Girshick, "Fast and accurate model scaling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 924–932.

[325] A. Ruiz and J. Verbeek, "Anytime inference with distilled hierarchical neural ensembles," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 11, 2021, pp. 1–9.
[326] B. Acun, M. Murphy, X. Wang, J. Nie, C.-J. Wu, and K. Hazelwood, "Understanding training efficiency of deep learning recommendation models at scale," in *Proc. IEEE Int. Symp. High-Performance Comput. Archit. (HPCA)*, Feb. 2021, pp. 802–814.
[327] H. D. Dixit, S. Pendharkar, M. Beadon, C. Mason, T. Chakravarthy, B. Muthiah, and S. Sankar, "Silent data corruptions at scale," 2021, *arXiv:2102.11245*.
[328] Y. Zhao, L. Wang, Y. Tian, R. Fonseca, and T. Guo, "Few-shot neural architecture search," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12707–12718.
[329] D. Wang, M. Li, C. Gong, and V. Chandra, "AttentiveNAS: Improving neural architecture search via attentive sampling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6418–6427.
[330] B. Zhang, R. Rajan, L. Pineda, N. Lambert, A. Biedenkapp, K. Chua, F. Hutter, and R. Calandra, "On the importance of hyperparameter optimization for model-based reinforcement learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 4015–4023.
[331] A. Zhang, R. McAllister, R. Calandra, Y. Gal, and S. Levine, "Learning invariant representations for reinforcement learning without reconstruction," 2020, *arXiv:2006.10742*.
[332] M. Panchenko, R. Auler, L. Sakka, and G. Ottoni, "Lightning BOLT: Powerful, fast, and scalable binary optimization," in *Proc. 30th ACM SIGPLAN Int. Conf. Compiler Construct.*, Mar. 2021, pp. 119–130.
[333] T. Li, A. Beirami, M. Sanjabi, and V. Smith, "Tilted empirical risk minimization," 2020, *arXiv:2007.01162*.
[334] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 6357–6368.
[335] K. Maeng, S. Bharuka, I. Gao, M. Jeffrey, V. Saraph, B.-Y. Su, C. Trippel, J. Yang, M. Rabbat, B. Lucia, and C. J. Wu, "Understanding and improving failure tolerant training for deep learning recommendation with partial recovery," *Proc. Mach. Learn. Syst.*, vol. 3, pp. 1–15, Mar. 2021.
[336] F. Lin, B. Bolla, E. Pinkham, N. Kodner, D. Moore, A. Desai, and S. Sankar, "Near-realtime server reboot monitoring and root cause analysis in a large-scale system," in *Proc. 51st Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw.-Supplemental (DSN-S)*, 2021, pp. 37–40, doi: 10.1109/DSN-S52858.2021.00027.
[337] Y. Xia, Y. Zhang, Z. Zhong, G. Yan, C. Lim, S. S. Ahuja, S. Bali, A. Nikolaidis, K. Ghobadi, and M. Ghobadi, "A social network under social distancing: Risk-driven backbone management during COVID-19 and beyond," in *Proc. NSDI*, 2021, pp. 217–231.
[338] E. J. Oughton, W. Lehr, K. Katsaros, I. Selinis, D. Bubley, and J. Kusuma, "Revisiting wireless internet connectivity: 5G vs Wi-Fi 6," *Telecommun. Policy*, vol. 45, no. 5, Jun. 2021, Art. no. 102127.
[339] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen, "A time-domain convolutional recurrent network for packet loss concealment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7148–7152.
[340] B. Reagen, W.-S. Choi, Y. Ko, V. T. Lee, H.-H.-S. Lee, G.-Y. Wei, and D. Brooks, "Cheetah: Optimizing and accelerating homomorphic encryption for private inference," in *Proc. IEEE Int. Symp. High-Performance Comput. Archit. (HPCA)*, Feb. 2021, pp. 26–39.
[341] M. Cowan, D. Dangwal, A. Alaghi, C. Trippel, V. T. Lee, and B. Reagen, "Porcupine: A synthesizing compiler for vectorized homomorphic encryption," in *Proc. 42nd ACM SIGPLAN Int. Conf. Program. Lang. Design Implement.*, Jun. 2021, pp. 375–389.
[342] S. Belkhale, R. Li, G. Kahn, R. McAllister, R. Calandra, and S. Levine, "Model-based meta-reinforcement learning for flight with suspended payloads," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1471–1478, Apr. 2021.
[343] J. C. Shih, F. Meier, and A. Rai, "A framework for online updates to safe sets for uncertain dynamics," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 5994–6001.
[344] E. Giraudy, P. Maas, S. Iyer, Z. Almquist, J. W. Schneider, and A. Dow, "Measuring long-term displacement using Facebook data," IDMC Global Rep. Internal Displacement (GRID), Geneva, Switzerland, Tech. Rep., 2021.
[345] J. Gray, A. Lerer, A. Bakhtin, and N. Brown, "Human-level performance in no-press diplomacy via equilibrium search," 2020, *arXiv:2010.02923*.
[346] V. Ha-Thuc, M. Wood, Y. Liu, and J. Sundaresan, "From producer success to retention: A new role of search and recommendation systems on marketplaces," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 2629–2630.
[347] S. Blackshear, K. Chalkias, P. Chatzigiannis, R. Faizullabhoy, I. Khaburzaniya, E. K. Kogias, J. Lind, D. Wong, and T. Zakian, "Reactive key-loss protection in blockchains," *IACR Cryptol. ePrint Arch.*, vol. 2021, p. 289, Mar. 2021.
[348] V. Conitzer, C. Kroer, E. Sodomka, and N. E. Stier-Moses, "Multiplicative pacing equilibria in auction markets," 2017, *arXiv:1706.07151*.
[349] J. Onaolapo, N. Leontiadis, D. Magka, and G. Stringhini, "SocialHEIST-ing: Understanding stolen Facebook accounts," in *Proc. 30th USENIX Secur. Symp. (USENIX Security)*, 2021, pp. 1–18.
[350] M. Bailey, P. Farrell, T. Kuchler, and J. Stroebel, "Social connectedness in urban areas," *J. Urban Econ.*, vol. 118, Jul. 2020, Art. no. 103264.
[351] M. Luria and N. Foulds, "Hashtag-forget: Using social media ephemerality to support evolving identities," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–5.
[352] V. Do, J. Atif, J. Lang, and N. Usunier, "Online selection of diverse committees," 2021, *arXiv:2105.09295*.
[353] F. de Lima Alcantara, "Diverse and inclusive representation in online advertising: An exploration of the current landscape and people's expectations," Meta, CA, USA, Tech. Rep., Mar. 2021. Accessed: Nov. 1, 2021. [Online]. Available: https://research.facebook.com/publications/diverse-and-inclusive-representation-in-online-advertising-an-exploration-of-the-current-landscape-and-peoples-expectations/
[354] V. Avadhanula, R. C. Baldeschi, S. Leonardi, K. A. Sankararaman, and O. Schrijvers, "Stochastic bandits for multi-platform budget optimization in online advertising," in *Proc. Web Conf.*, Apr. 2021, pp. 2805–2817.
[355] D. Sinha, K. A. Sankararaman, A. Kazerouni, and V. Avadhanula, "Multi-armed bandits with cost subsidy," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 3016–3024.
[356] J. Truong, S. Chernova, and D. Batra, "Bi-directional domain adaptation for Sim2Real transfer of embodied navigation agents," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 2634–2641, Apr. 2021.
[357] *Detectron2 GitHub*. Accessed: Nov. 11, 2021. [Online]. Available: https://github.com/facebookresearch/detectron2

SANG-MIN PARK received the Ph.D. degree in computer engineering from the Department of Computer Science and Engineering, Korea University, in 2016. His current research interests include natural language processing, cognitive science, sentiment analysis, causal inference, multi-modal analysis, personalized service, generative model, and reinforcement learning.

YOUNG-GAB KIM (Member, IEEE) received the B.S. degree in biotechnology and genetic engineering (minor in computer science and engineering) and the M.S. and Ph.D. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 2001, 2003, and 2006, respectively. He was an Assistant Professor with the School of Information Technology, Catholic University of Daegu. He is currently an Associate Professor with the Department of Computer and Information Security and Convergence Engineering for Intelligent Drone, Sejong University. As a Korean ISO/IECJTC1 Member, he has contributed to the development of data exchange standards. He has published more than 190 research articles in the fields of computer science and information security. His current research interests include big data security, network security, home networks, security risk analysis, and security engineering.