Guiding AI-Generated Digital Content with Wireless Perception

Jiacheng Wang, Hongyang Du, Dusit Niyato, Zehui Xiong, Jiawen Kang, Shiwen Mao, and Xuemin (Sherman) Shen

Abstract

Advancements in artificial intelligence (AI), and the surge in diverse training data, have facilitated AI generated content (AIGC). Despite high efficiency, the inherent instability of AI models poses challenges in creating user-specific content, especially when creating an avatar for a user. To address this issue, this article integrates wireless perception (WP) with AIGC and introduces WP-AIGC, a unified framework that leverages a user skeleton obtained by WP to guide AIGC, thereby generating the avatar that aligns with the user's actual posture. Specifically, WP-AIGC first employs a novel multi-scale perception technology to sense posture in the physical world and construct the user skeleton. Then, the skeleton and the user's requirements are conveyed to the AIGC, thereby guiding the creation of the avatar. Furthermore, WP-AIGC can adjust the computing resources allocated to perception and AIGC based on user feedback, thereby optimizing the service. Experimental results verify the effectiveness of the service. With limited computing resources, WP-AIGC achieves optimal QoS of 3.75 when four links are involved in perception.

Introduction

The spectacular growth of various types of data, hardware upgrades, and the advancement of artificial intelligence (AI) models has led to the emergence of AI generated content (AIGC), which can imitate human behavior to create digital content [1]. Specifically, AIGC refers to the AI-enabled methods (which are able to automatically produce, manipulate, and modify multi-modal digital content) and the corresponding generated content [2]. Due to the ability of automatically producing various kinds of high-quality digital content, the AIGC is gaining increasing attention, especially with the rapid integration of the physical world and virtual digital world.

At the function and application levels, AIGC enables autonomous content creation through AI [3], and boosts the development of various applications. Taking the virtual interactive game in Metaverse as an example, AIGC can generate avatars and create the corresponding scenarios according to users' requirements, thereby constructing a complete virtual world for users to explore. During this process, the user's needs, for example, prompts, can be transmitted to the AIGC model through various ways such as voice and text [4]. Yet, some information is challenging to convey through words, such as the user's posture in the physical world. Some feasible methods include utilizing cameras (such as Kinect — https://learn.microsoft.com/en-us/windows/apps/design/devices/kinect-for-windows) or on-body sensors (such as Sony MO-COPI — https://www.sony.net/Products/mocopi-dev/en) to capture the user's posture, which can be combined with user's prompts and then fed into an AIGC model to produce digital content. For instance, systems like Sony MO-COPI typically use high-quality sensor such as accelerators and gyroscopes to gather movement data and construct the user's posture. These systems are known for their reliable and impressive performance. However, the specialized sensors can be costly, and wearing them for extended periods might be uncomfortable to users.

Besides the above mentioned methods, wireless perception is also a suitable method for providing AIGC perception support [5]. Compared to the aforementioned camera and sensor-based systems, wireless perception operates in a passive manner, without requiring users to wear or carry any devices, making it more user-friendly and easily acceptable. Additionally, it can be achieved by leveraging existing wireless communication signals without necessitating any other specialized devices. Although it may be slightly inferior in terms of stability and accuracy, wireless perception offers lower costs and a wider range of applicability, hence presenting broad prospects for future applications.

Building on the above discussion, we present the first wireless perception based AIGC (WP-AIGC) framework, which seamlessly integrates wireless perception with AIGC, affording users with better digital content generation services. Taking the virtual game as an example, the proposed framework is illustrated in Fig. 1. As can be seen, in the physical world, the wireless signals are used to sense the user and predict the corresponding skeleton. Subsequently, the AIGC model generates the avatar according to the user's prompts and the obtained skeleton. In WP-AIGC, the number of sensors affects the perception accuracy, that is, the similarity between the predicted skeleton and the user's actual posture, while the number of infer-

Jiacheng Wang, Hongyang Du and Dusit Niyato are with Nanyang Technological University, Singapore; Zehui Xiong is with Singapore University of Technology and Design, Singapore; Jiawen Kang (corresponding author) is with Guangdong University of Technology, China; Shiwen Mao is with Auburn University, USA; Xuemin Shen is with the University of Waterloo, Canada.

Digital Object Identifier: 10.1109/MWC.008.2300162

1536-1284/24/$25.00 © 2024 IEEE

As can be seen, in the physical world, the wireless signals are used to sense the user and predict the corresponding skeleton. Subsequently, the AIGC model generates the avatar according to the user's prompts and the obtained skeleton

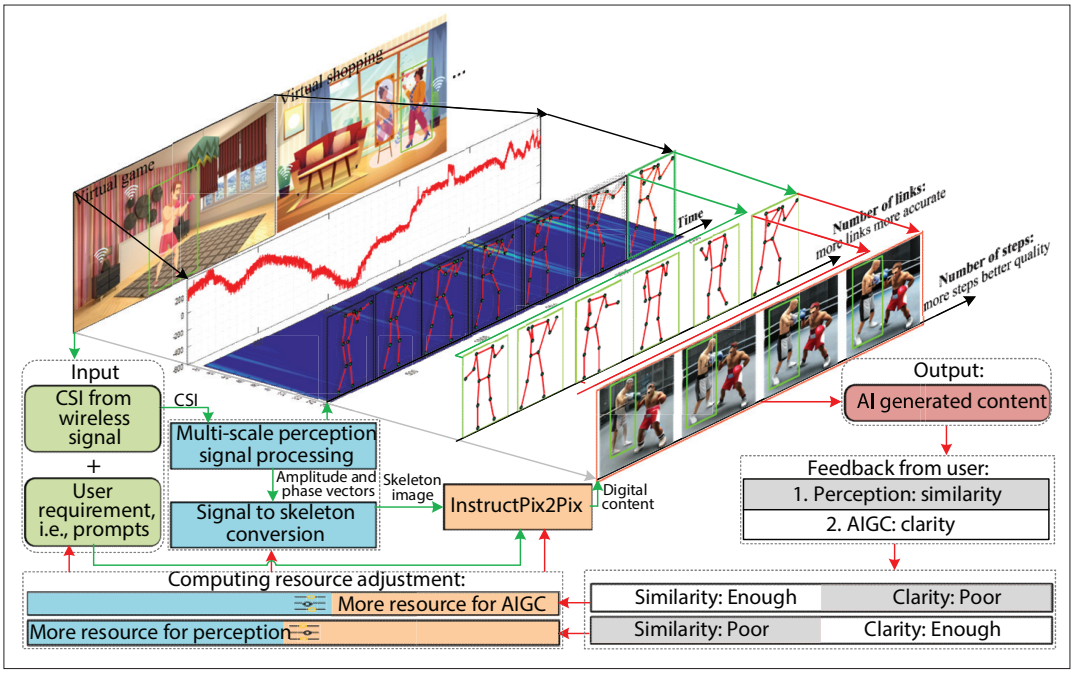

FIGURE 1. The overall structure of WP-AIGC. During the operation process, the channel state information (CSI) extracted from wireless signal is first processed, and then a pre-trained network is used to predict the skeleton based on the processed CSI. After that, the obtained skeleton and the user's requirements (e.g., prompts) are passed to the AIGC model to generate the corresponding avatar, which is finally presented to the user via devices like virtual reality headset. Furthermore, users can provide feedback to WP-AIGC, which then optimizes the computing resources based on the feedback, thereby improving the overall quality of service (QoS).

ence steps in AIGC model affects the quality of generated avatar. To balance the perception accuracy and content quality, WP-AIGC receives the user's feedback on satisfaction with generated content and adjusts the computing resources at the server allocated to perception and AIGC accordingly, thereby enhancing user experience. Overall, the main contributions of this article are as follows:

- We propose WP-AIGC, the first framework integrating wireless perception and AIGC for providing avatar generation services to users. The WP-AIGC includes wireless perception, AIGC based content generation, and a feedback interface, which receives the user's feedback and optimizes computing resources at the edge server to enhance the user experience.
- We propose the multi-scale wireless perception, which refers to perform large-scale and small-scale perception on users in sequence. Unlike existing works, the proposed method allows for perceptions at different scales to assist each other by sharing the results, thereby enhancing the overall perceptual performance.
- Leveraging the collected wireless perception data and the AIGC model, we provide a practical and compelling use case to verify the feasibility of the proposed framework, lighting the way of providing virtual service via the combination of wireless perception and AIGC.

AIGC and Wireless Perception Technologies

In this section, we provide a comprehensive review of AIGC and wireless perception technologies.

AIGC

The development of AIGC can be divided into three main stages. The initial stage (around 1950s to 1990s) is characterized by limited technologies, and only small-scale experiments are feasible, resulting in products such as the "Illiac Suite" [6]. In the second stage (around 1990s to 2010s), significant advances in deep learning, the evolution of internet services, coupled with the accumulation of user data pushed AIGC into real-world applications. We are now in the third stage, with the emergence of new and better AI models leading to the more powerful and intelligent AIGC, capable of producing a variety of digital content that imitates or even exceeds human creations.

Basic to advanced, AIGC functions include content repair, enhance, edit, and generate. Specifically, repair refers to fix missing content, such as using diffusion models [7] to recover missing pixels in an image. Besides that, AIGC can also enhance the image quality by increasing the contrast, improving pixel quality and clarity, as shown in Fig. 2. The content editing and generation are more advanced features of AIGC. Concretely, editing means that AIGC is capable of conducting operations, such as modifying and replacing, on the specified content. Such a function can be used to handle sensitive content, such as replacing sensitive people or things without changing the overall layout and style of the image [8]. Content generation is the function of AIGC that distinguishes it from other AI models. Benefiting from a large amount of training data, algorithmic progresses and hardware upgrades, AIGC can now generate not only images, but also videos, codes, and manuscripts, which can greatly improve productivity in creating digital content. According to the above described definition and functions, several features of AIGC can be summarized as follows.

Automatic: Given a specified task or order, AIGC can automatically produce digital content and present it in various forms, such as pictures and videos, which is more productive than tradi-

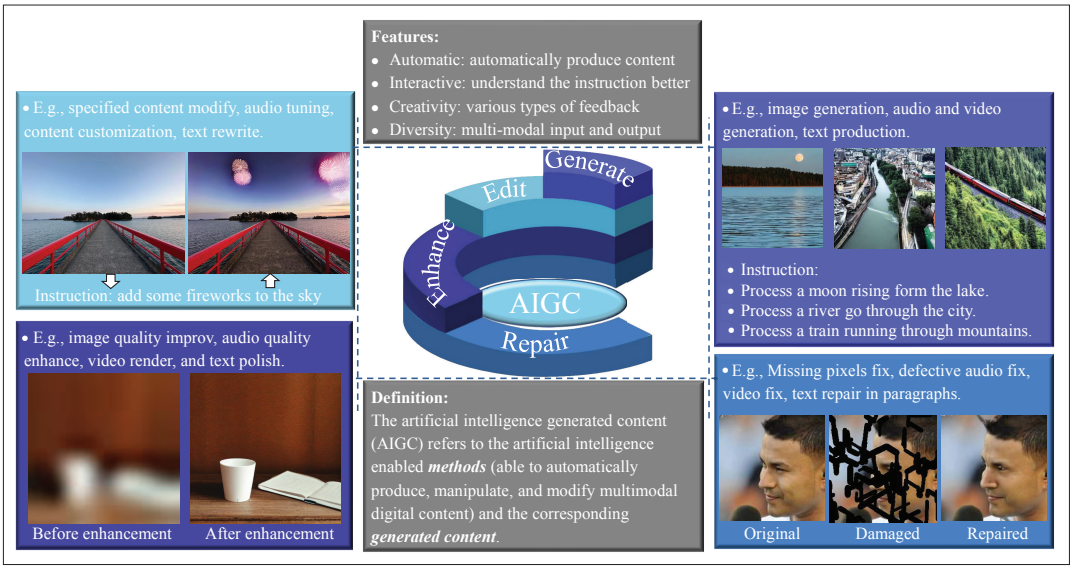

FIGURE 2. The definition, features, and the corresponding applications of AIGC.

tional ways that often require human involvement, such as professionally generated content (PGC) and user generated content (UGC).

Interactive: Due to the massive training data and human-like way of thinking, AIGC can better understand the user's thoughts or instructions through prompts, making AIGC can interact with people in a more natural way than other AI models. Such a feature is expected to further improve the interaction experience between humans and machines.

Creativity: Different from the traditional AI model with limited output space in most cases (such as a classifier), AIGC has diverse answers to the same question, demonstrating its creative ability that can promote the diversification of digital content.

Diversity: AIGC not only supports multi-modal input, but the generated digital content can also be presented in different forms, which is another important characteristics that distinguishes AIGC from other AI models. As a result, AIGC can assist in a variety of digital content production in different fields.

In Fig. 2, the definition, features and some applications of AIGC are presented. It is clear that AIGC would significantly improve the content and information production efficiency in the near future, thereby revolutionizing the traditional digital content production and consumption mode.

Wireless Perception

With the rise of ISAC technologies, using wireless signals for perception gains considerable traction. The essence of this approach is that environmental characteristics affect wireless signal propagation, thereby encoding the signal with information about its surroundings [9]. By applying signal processing methods, we can extract these characteristics, so as to realize the perception of the physical environment. According to the principle, we can observe that wireless perception has two notable advantages:

- The first is the wide coverage. Theoretically, wireless perception can be applied anywhere covered by wireless signals. It should be noted that wireless signals can vary under different scenarios, which may result in different perception abilities.
- The second is that users are not required to wear any device during the perception process, eliminating the discomfort associated with prolonged device wear and the concern for device recharging.

So far, researchers have conducted extensive studies on wireless perception, covering human detection, localization, and recognition, such as posture and gesture. For instance, user detection refers to the determination of the presence or absence of a user in a specific area by analyzing signal fluctuations. The main steps often include signal denoising, feature extraction, and user detection. Another significant area of research is the reconstruction of user posture. In [5], the authors first extract the amplitude and phase of the CSI, captured from multiple sensing links into images, and then use the neural networks to predict the user skeleton. Similarly, in WiSPPN [10], the authors propose a fully convolutional network, which consists of an encoder, feature extractor, and decoder, to predict the pose adjacent matrix. This matrix can then be utilized to reconstruct the user's posture. While these approaches to posture reconstruction can yield good results, they often overlook the impact of user's location and posture on different perception links, lacking a specific processing for each one. Therefore, they struggle to fully extract and utilize the user-related information within the CSI measurement.

Based on foregoing discussion, particularly the role of AIGC, the principle and advantages of wireless perception, it is natural to consider using the output of wireless perception, that is, user skeleton, as the input of AIGC, thereby generating virtual avatars that are closely align with the user's actual posture. Yet, integrating these two technologies poses important technical challenges, which are discussed next.

The Proposed Framework

This section presents the proposed WP-AIGC framework, including research challenges and implementation processes.

So far, researchers have conducted extensive studies on wireless perception, covering human detection, localization, and recognition, such as posture and gesture.

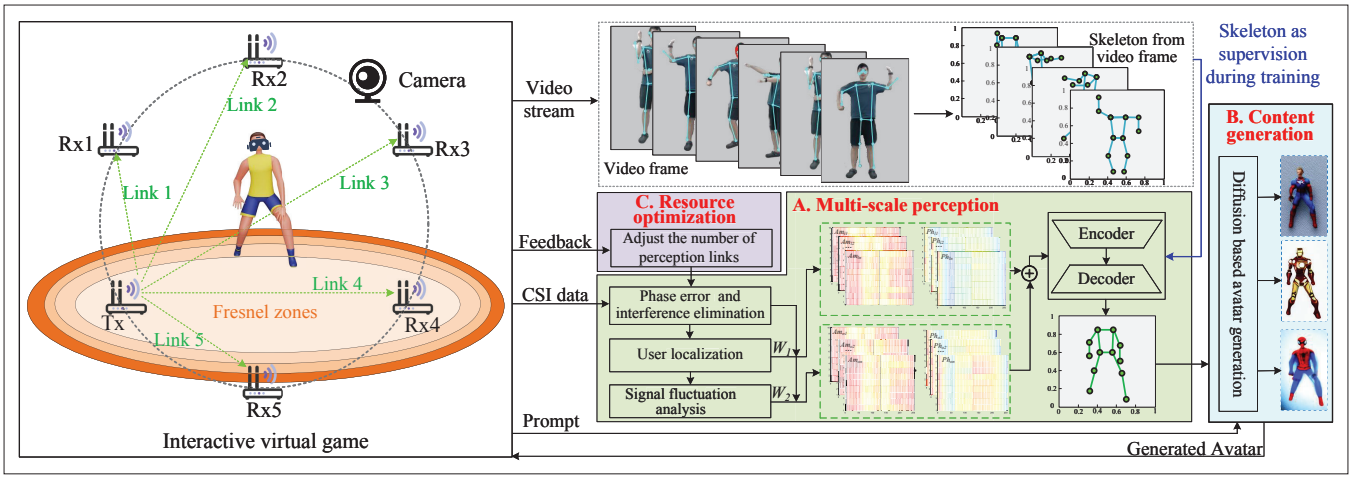

FIGURE 3. The three main parts of the proposed WP-AIGC. Here, Tx is the wireless signal transmitter, and Rx is the receiver. The green dashed lines are the transmission links and the orange concentric circles represent the Fresnel zone formed by Tx and Rx4. Note that during the training, the human skeleton extracted from video frames is used as the ground truth to supervise the predicted skeleton, so as to complete the training. When providing services, the trained network can directly convert the obtained CSI into a skeleton without the camera, thus constituting a device-free solution.

Research Challenges

Multi-Scale Wireless Perception: When guiding AIGC with wireless perception, the perception of users at various scales in the physical space is essential. For instance, in Fig. 1, the user's location in the physical world corresponds to the virtual character's position in the boxing ring, while posture facilitates the virtual avatar construction. Therefore, multi-scale perception is vital for WP-AIGC. This perception task necessitates obtaining the user's location and posture from the same data set. Meanwhile, perception at different scales needs to mutually enhance each other, which poses a considerable challenge.

Computing Resource Optimization: Balancing computing resources between perception and AIGC is critical. With limited resources, if less is allocated to perception, more can support AIGC, which fosters content generation but may weaken the perception. Conversely, allocating more to perception may leave AIGC with insufficient resources. This ensures accurate perception, but the generated content may not meet user needs due to resource constraints. In practice, users have different preferences for perception and AIGC, and therefore effective resource balance depends on the user feedback.

Digital Content Generation: Generative diffusion models offer a balance between tractability and flexibility, aiming to bridge the gap between simple, analytically-evaluable models and generative model that can capture intricate data structures. While diffusion models can efficiently model complex data distributions, their primary challenge lies in the generation process. Specifically, they necessitate an extended sequence of Markovian diffusion steps for producing samples. This multi-step approach, intrinsic to diffusion models, imposes computational demands. As a consequence, time and resources required for sample generation may render diffusion models less optimal for applications where rapid response or limited computational power is paramount.

Content Quality Assessment: The evaluation of content quality is not only directly related to the QoS of the framework, but also can guide the allocation of computing resources. The content quality evaluation should consider two major aspects. The first is the accuracy of the generated content, which depends on the wireless perception accuracy and whether the user's requirements are accurately transmitted to AIGC. The second aspect is the quality of content presentation, such as whether images are clear or videos are smooth. The evaluation of the first aspect depends more on user feedback, while the second aspect can be analyzed through numerical analysis.

The Proposed WP-AIGC

As illustrated in Fig. 3, the proposed framework comprises three parts. The first is multi-scale perception, which combines signal processing with machine learning to transform CSI into a user skeleton. The second part involves content generation, producing appropriate digital content based on user's needs and the obtained skeleton. The third part adjusts the resource allocation based on user feedback, ensuring a balanced performance between perception and AIGC.

Multi-Scale Sensing: To achieve a desired perception accuracy, we propose to perceive users in an order of large-scale to small-scale, as shown in Fig. 3. Specifically, during perception, a transmitter (Tx) sends wireless signals, and the receiver (Rx) uses multiple antennas to capture the signals and extract CSI. Note that, at the Rx, one antenna is used to receive the reference signal, while the rest form an array to capture the surveillance signals. Then, to eliminate the direct signals and static object induced reflections, which lack valuable information about the user, we apply the fast Fourier transform to the CSI, nullify the zero-frequency components, and then convert the processed CSI to the time domain via the Inverse fast Fourier transform.

Based on data converted back to the time domain, we complete multi-scale perception through the following steps:

1. The matrix pencil algorithm is used to estimate the angle of arrival (AoA) and time of flight (ToF) of the user induced reflection. Using estimated parameters, the user's location is calculated.
2. The reciprocal of distance between the user and perception link is used to weight the CSI amplitude and phase of each perception link.

3. The weighted CSI amplitudes and phases are summed respectively to obtain the amplitude and phase vectors to complete large-scale perception.
4. Based on estimated user location and the Fresnel zone principle,[1] the unbiased variance of each link is calculated, which indicates the fluctuation characteristics of the perception link.
5. The CSI amplitude and phase from each perception link are weighted by the unbiased variance, then summed to produce another pair of amplitude and phase vectors, enabling small-scale perception.
6. The amplitude and phase vectors obtained by steps 3 and 5 are summed, respectively, to obtain the aggregated amplitude and phase vectors, thereby completing multi-scale perception.

The aggregated vectors derived here serve as the input for the network that converts signals into a user skeleton, which will be elaborated in the next section.

Signal to Skeleton Conversion: Using the obtained aggregated vector, a neural network, which contains an encoder and a decoder, is jointly trained to map the extracted amplitude and phase vectors to human skeleton images. Here, the encoder employs strided convolutional networks to extract information from aggregated vectors, and the decoder applies resize convolutions to reconstruct the user pose. Specifically, the encoder network consists of three layers of 3 × 3 convolutions with 2 × 2 strides, each followed by a 1 × 1 convolutional layer with a stride of 1 × 1. The rectified linear unit activation functions are applied after each layer and a fully connected layer is utilized after the final convolutional layer to convert images directly. The decoder network consists of the total of seven layers, where the first two layers use 1 × 1 kernels with a stride of 1 × 1, and the subsequent 5 layers use 3 × 3 convolutions with a stride of 1 × 1.

To ensure skeleton prediction, we synchronize CSI collection with video capture and extract skeleton from video stream for training supervision, as shown in Fig. 3. Assuming [**V**, **A**, **P**] is a synchronized data set, where **V** is the user pose extracted from the video, and **A** and **P** correspond to the aggregated amplitude and phase vectors, respectively. Then, the network takes **A** and **P** as the inputs to predict the user skeleton, which is optimized through the supervision of **V**. Here, the training objective is to minimize the difference between the predicted skeleton and the corresponding **V**, with the loss function defined as the average Euclidean distance error. The network is implemented using TensorFlow and trained over 64 epochs with a batch size of 32, with a learning rate of 0.001.

Digital Content Generation: After obtaining the user skeleton, the InstructPix2Pix [12], which can edit images based on user's instructions, is utilized to generate corresponding virtual digital content. The construction of InstructPix2Pix consists of two steps: generating training data and training the model.

In the first step, a fine-tuned GPT-3[2] is used to generate instructions and edited captions, and then StableDiffusion[3] is combined with Prompt-to-Prompt[4]Nto generate paired images based on the paired instructions and captions.

In the second step, using the generated pairs of images and corresponding instructions, a conditional diffusion model is trained to predict noise added to the given image and the text instruction.

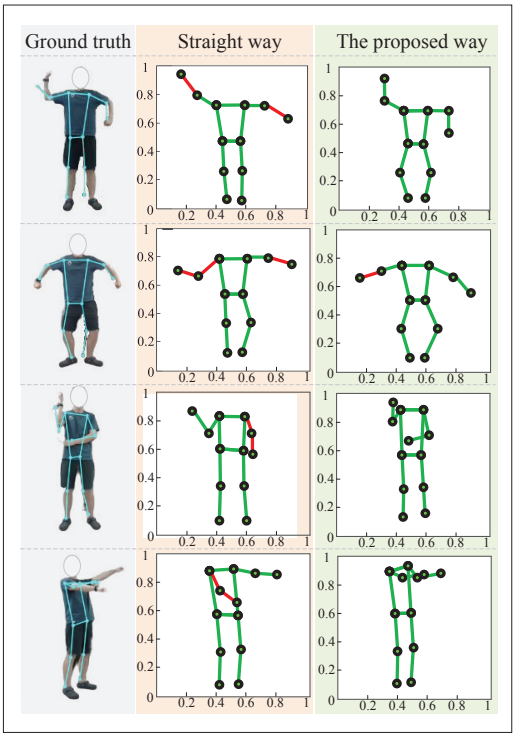

FIGURE 4. Comparison of experimental results of different perception methods.

During the training process, the available weights of the InstructPix2Pix model are initialized with a pre-trained StableDiffusion checkpoint. Meanwhile, to enable the trained model to perform conditional or unconditional denoising with respect to either or both conditional inputs, 5 percent of the sample images, 5 percent of the sample instructions, and 5 percent of the sample images with instructions are randomly set to an empty set. Based on such capability, two scale guidance parameters are further introduced to trade off how strongly the generated samples correspond with the input image and how strongly they correspond with the edit instruction. After training, WP-AIGC utilizes this network to generate digital content, with the obtained user skeleton as the input and user's service requirements as editing instructions.

Computing Resource Adjustment: In an initial stage of operation, WP-AIGC first derives an optimal resource allocation strategy to maximize QoS, based on a mapping relationship between computing resources and the TV, BRISQUE, and similarity.[5] Here, the QoS is calculated as the sum of the inverse of both normalized Total Variation (TV) and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE), and the similarity. WP-AIGC then allocates computing resources according to this strategy to deliver services, that is, generate the corresponding avatar and display it to the user via devices like virtual reality headset. After that, WP-AIGC monitors user feedback and adjusts its resource allocation strategy accordingly.

For instance, in an interactive boxing game, if a user throws a punch with the right hand and the avatar fails to mimic this motion due to insufficient perceptual accuracy, then the user can clearly observe this discrepancy, being unsatisfactory, and report to the system. In response, WP-AIGC

Using the obtained aggregated vector, a neural network, which contains an encoder and a decoder, is jointly trained to map the extracted amplitude and phase vectors to human skeleton images

[1] The Fresnel zone consists of a series of concentric ellipses. When an object moves perpendicularly to the elliptical boundary, it passes through more ellipses, causing greater fluctuations in the received signal. Conversely, when it moves parallel to the boundary, the signal fluctuations are smaller [11].

[2] Fine-tuning GPT-3 involves two steps: generating edit instructions and modified captions via a large language model and then fine-tuning GPT-3 with a dataset containing input captions, edit instructions, and output captions.

[3] StableDiffusion is a generative model that synthesizes high-resolution images efficiently by applying diffusion models within the latent space of autoencoders, enhanced by cross-attention layers, enabling diverse image creation tasks with reduced computational demands [13].

[4] Prompt-to-Prompt [14] is a text-driven image editing framework that manipulates cross-attention layers to alter images based on textual prompt modifications without direct pixel space adjustments.

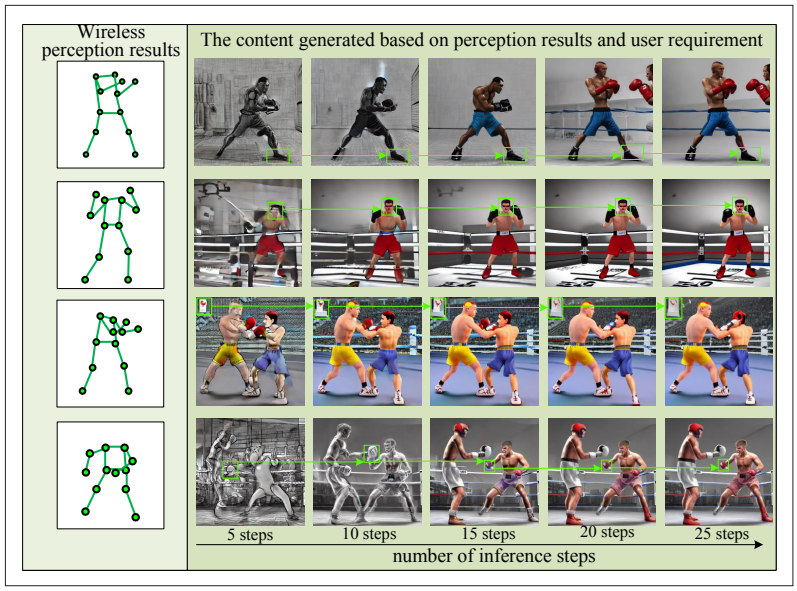

FIGURE 5. The impact of the AIGC's inference steps on the avatar generation. The image on the left is the input skeleton. The images on the right are the corresponding virtual avatars generated based on different seeds. As indicated by the details marked in green, more inference steps results in avatars with higher quality.

allocates more resources to perception (with the remaining resources dedicated to AIGC) to activate an additional perception link in the service, thereby enhancing the perception accuracy. Similarly, if the feedback indicates inadequate quality of the virtual avatar, WP-AIGC reduces the number of perception links, freeing up more resources for AIGC to enhance the content generation.

Case Study

To validate WP-AIGC, we conducted tests in a practical scenario. Specifically, we build a 802.11ac protocol based wireless perception platform, which contains a signal transmitter and five receivers for signal transmission and CSI extraction. The extracted CSI is processed offline on a server to complete multi-scale perception, user skeleton generation, and virtual avatar generation. The server is built on the standard Ubuntu 20.04 system, equipped with an AMD Ryzen Threadripper PRO 3975WX 32-core processor and an NVIDIA RTX A5000 Graphics Processing Unit (GPU).

To begin with, we compare the skeletons generated using our multi-scale perception technology with those produced without it, where skeletons are depicted by joint points. As illustrated in Fig. 4, although the overall similarity between the skeletons generated by the two methods and the real human pose is close, there are discernible differences in details, highlighted by the red-marked areas in the images. The differences are particularly noticeable when the arms are positioned close to the torso, as shown in the third row of images, where signals reflected by the torso and arms are difficult to distinguish. In such scenarios, the multi-scale perception approach yields a more accurate skeleton. The reason is that the proposed method weights the links that carry more information about the user to play a greater role in the skeleton generation process, thereby improving the avatar construction accuracy.

After the validation of multi-scale perception, we conduct an analysis on avatar generation, using boxing as an example. The results are presented in Fig. 5. From the results, we can see that InstructPix2Pix is capable of generating virtual avatars that mirror the user's posture based on the obtained skeleton and user's instruction, validating the effectiveness of the proposed WP-AIGC. Besides that, it can be observed that, using the identical skeleton and instruction set, an increase in the number of inference steps leads to an improvement in the avatar quality. This improvement is evident in two key aspects: enhanced clarity, as indicated by the avatar's facial features in the second row; and better color matching, as demonstrated by the boxing gloves of the avatar presented in the fourth row. Therefore, when sufficient computing resources are allocated to AIGC, the quality of the generated avatar can be enhanced by appropriately increasing the number of inference steps.

However, in practice, available computing resources are often limited. Here, based on the server used, processing CSI data from a single perception link takes 0.097 seconds, generating a user skeleton takes 0.048 seconds, and performing a single inference step in AIGC takes approximately 0.05 seconds. Therefore, given the above mentioned parameters, we analyze the relationship among the number of perception links, perception accuracy (measured by the similarity between the skeleton and the actual posture of the user), and avatar quality (measured by TV and BRISQUE), under a content refresh rate constraint of 1 Hz. The results are shown in Fig. 6.

The experimental results show that increasing the number of perception links improves the perception accuracy, that is, the similarity. However, this reduces the computing resources left for AIGC, resulting in an increase in TV and BRISQUE, that is, a decrease in image quality. For example, as the number of perception links increases from 4 to 5, the similarity improves from 0.86 to 0.95. However, at the same time, the BRISQUE rises from 32.05 to 46.30, and the TV increases from 79.32 to 149.3. Based on the relationships in Fig. 6, it can be calculated that WP-AIGC attains its maximum QoS of 3.75 when four perception links are involved, which provides a reference for the initial allocation of resources. Given that user's demands for perceptual accuracy and image quality may vary across different applications, such results show that it is crucial to obtain user feedback and adjust computing resources in situations where resources are limited.

Future Directions

AIGC Model Selection Strategy

Nowadays, various models are available for digital content generation, each of them excels in different domains and consumes different amounts of resource. Therefore, it is crucial to design a model selection mechanism that considers user requirements, resource availability, and other relevant factors. During the design, it is necessary to analyze the user's preferences when historical data is available, thereby enabling a customized AIGC experience. At the same time, given the long generation time required by common AIGC models like StableDiffusion in digital content generation, the selection mechanism needs detailed consideration of latency issues, as these directly impact the user experience.

[5] The TV characterizes the smoothness of an image, while BRISQUE quantifies the potential loss of naturalness in an image.

Edge Computing Resource Optimization

When deploying AIGC in mobile networks, optimizing edge computing resources is essential. The resources here include computing resources required for digital content generation, storage capacity, and transmission resources consumed for content delivery, such as bandwidth and transmission power [2]. The design process for the joint resource optimization model must ensure the quality of content and minimize latency, while also preventing overloading the AIGC model due to an excess of tasks. Additionally, the impact on other network services must be taken into account. One potential approach is to utilize deep reinforcement learning to allocate various network resources.

Guiding AIGC with Other Promising Techniques

Regarding the proposed WP-AIGC, further research can be conducted on how to expand the range that users can move around and the number of users, so as to improve the system's practicality. Besides, AIGC can be combined with other promising technologies, such as eye-tracking and brain-computer interfaces [15], to guide or even control AIGC in generating more complex digital content that better meets user's requirements. However, signals such as human brain waves are weaker, and are also influenced by human emotions and thoughts, making them contain more information. Therefore, when using brain waves to guide the AIGC, the difficulty lies in accurately inferring user needs based on raw signals and efficiently transmitting these needs to the AIGC model. Meanwhile, ensuring the security of both the original signal and the generated content is crucial, and blockchain can be introduced here to achieve these objectives.

Conclusions

In this article, we propose the WP-AIGC framework, which first leverages the multi-scale wireless perception technology to sense users in the physical world and predict the corresponding skeleton. It then sends the obtained skeleton to the AIGC model, which generates the digital avatar according to the user's prompts. Unlike most existing works that take descriptive text as input, WP-AIGC takes perception results as input. This can more accurately convey the user's posture in physical world to AIGC, thereby generating virtual avatars that are more closely aligned with the actual user. Furthermore, WP-AIGC can optimize the allocation of computing resources based on user feedback, thereby enhancing the QoS. To our knowledge, WP-AIGC is the first framework to combine wireless perception with AIGC technology to achieve digital content generation. In the future, we will investigate specific technical issues in the process of combining perception with AIGC, thereby further unleashing the productivity of AIGC.

Acknowledgment

This research is supported by the National Research Foundation, Singapore, and Infocomm Media Development Authority under its Future Communications Research & Development Programme, DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-019 and FCP-ASTAR-TG-2022-003), Energy Research Test-Bed and Industry Partnership Funding Initiative, Energy Grid (EG) 2.0 programme, DesCartes and the Campus for Research Excellence and Technological Enterprise (CREATE) programme, and MOE Tier 1 (RG87/22). The work is also supported by NSFC under grant No. 62102099, U22A2054, and No. 62101594, and the Pearl River Talent Recruitment Program under Grant 2021QN02S643, and Guangzhou Basic Research Program under Grant 2023A04J1699. S. Mao's work is supported in part by the NSF Grant CNS-2148382.

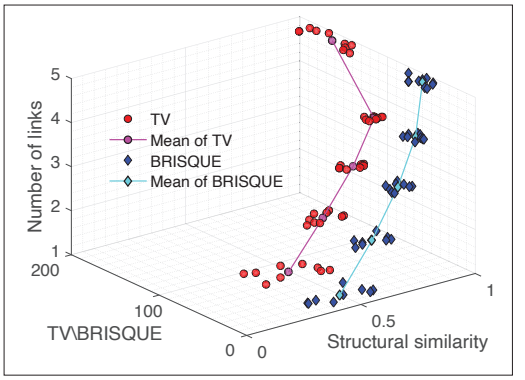

FIGURE 6. The relationship among the number of perception links, similarity, TV, and BRISQUE.

Regarding the proposed WP-AIGC, further research can be conducted on how to expand the range that users can move around and the number of users, so as to improve the system's practicality.

References

[1] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," *Proc. 2018 IEEE Int'l. Workshop on Information Forensics and Security*, IEEE, 2018, pp. 1–7.
[2] H. Du *et al.*, "Enabling AI-Generated Content (AIGC) Services in Wireless Edge Networks," arXiv preprint arXiv:2301.03220, 2023.
[3] A. Lugmayr *et al.*, "Repaint: Inpainting Using Denoising Diffusion Probabilistic Models," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 11,461–71.
[4] H. H. Thorp, "ChatGPT is Fun, but Not an Author," *Science*, vol. 379, no. 6630, 2023, p. 313.
[5] L. Guo *et al.*, "From Signal to Image: Capturing Fine-Grained Human Poses With Commodity Wi-Fi," *IEEE Commun. Letters*, vol. 24, no. 4, 2019, pp. 802–06.
[6] L. A. Hiller Jr and L. M. Isaacson, "Musical Composition With a Highspeed Digital Computer," *J. Audio Engineering Society*, vol. 6, no. 3, 1958, pp. 154–60.
[7] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 8780–94.
[8] G. Harshvardhan *et al.*, "A Comprehensive Survey and Analysis of Generative Models in Machine Learning," *Computer Science Review*, vol. 38, 2020, p. 100285.
[9] H. Du *et al.*, "Semantic Communications for Wireless Sensing: Ris-Aided Encoding and Self-Supervised Decoding," *IEEE JSAC*, vol. 41, no. 8, 2023, pp. 2547–62.
[10] F. Wang *et al.*, "Can WiFi Estimate Person Pose?" arXiv preprint arXiv:1904.00277, 2019.
[11] H. Wang *et al.*, "Human Respiration Detection With Commodity Wifi Devices: Do User Location and Body Orientation Matter?" *Proc. 2016 ACM Int'l. Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 25–36.
[12] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to Follow Image Editing Instructions," arXiv preprint arXiv:2211.09800, 2022.
[13] R. Rombach *et al.*, "High-Resolution Image Synthesis With Latent Diffusion Models," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2022, pp. 10,684–95.
[14] A. Hertz *et al.*, "Prompt-to-Prompt Image Editing With Cross Attention Control," arXiv preprint arXiv:2208.01626, 2022.
[15] D. Wu *et al.*, "Transfer Learning for Eeg-Based Brain– Computer Interfaces: A Review of Progress Made Since 2016," *IEEE Trans. Cognitive and Developmental Systems*, vol. 14, no. 1, 2020, pp. 4–19.

Biographies

Jiacheng Wang (jiacheng.wang@ntu.edu.sg) is the postdoctoral research fellow in Computer Science and Engineering at Nanyang Technological University, Singapore. Prior to that, he received the Ph.D. degree in School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include wireless sensing, generative artificial intelligence, and semantic communications.

Hongyang Du (hongyang001@e.ntu.edu.sg) is working toward his Ph.D. degree with the School of Computer Science and Engineering, the Energy Research Institute @ NTU, Interdisciplinary Graduate Program, Nanyang Technological University, Singapore. He was the recipient of IEEE Daniel E. Noble Fellowship Award in 2022. His research interests include generative AI, semantic communications, and communication theory.

Dusit Niyato (dniyato@ntu.edu.sg) is a professor in the School of Computer Science and Engineering, at Nanyang Technological University, Singapore. He received Ph.D. in Electrical and Computer Engineering from the University of Manitoba, Canada in 2008. His research interests are in the areas of sustainability, edge intelligence, decentralized machine learning, and incentive mechanism design.

Zehui Xiong (zehui_xiong@sutd.edu.sg) is currently an Assistant Professor with the Pillar of Information Systems Technology and Design, Singapore University of Technology Design, Singapore. His research interests include wireless communications, network games and economics, blockchain, and edge intelligence.

Jiawen Kang (kavinkang@gdut.edu.cn) received the Ph.D. degree from the Guangdong University of Technology, China in 2018. He has been a postdoc at Nanyang Technological University, Singapore from 2018 to 2021. He currently is a full professor at Guangdong University of Technology, China. His research interests focus on blockchain, security and privacy protection.

Shiwen Mao (smao@ieee.org) received his Ph.D. in electrical and computer engineering from Polytechnic University, Brooklyn, NY. He is a Professor and Earle C. Williams Eminent Scholar, and Director of the Wireless Engineering Research and Education Center at Auburn University. His research interests include wireless networks and multimedia communications.

Xuemin (Sherman) Shen (sshen@uwaterloo.ca) received the PhD degree in electrical engineering from Rutgers University, New Jersey, in 1990. He is a University professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, network security, Internet of Things, 5G and beyond.