Enabling AI-Generated Content Services in Wireless Edge Networks

Hongyang Du, Zonghang Li, Dusit Niyato, Jiawen Kang, Zehui Xiong, Xuemin (Sherman) Shen, and Dong In Kim

Abstract

Artificial intelligence-generated content (AIGC) refers to the use of AI to automate the information creation process while fulfilling the personalized requirements of users. However, due to the instability of AIGC models — for example, the stochastic nature of diffusion models — the quality and accuracy of the generated content can vary significantly. In wireless edge networks, the transmission of incorrectly generated content may unnecessarily consume network resources. Thus, a dynamic AIGC service provider (ASP) selection scheme is required to enable users to connect to the most suited ASP, improving the users' satisfaction as well as the quality of generated content. In this article, we first review the AIGC techniques and their applications in wireless networks. We then present the AIGC-as-a-service (AaaS) concept and discuss the challenges in deploying AaaS at the edge networks. It is essential to have performance metrics to evaluate the accuracy of AIGC services. Thus, we introduce several image-based perceived quality evaluation metrics. Then, we propose a general and effective model to illustrate the relationship between computational resources and user-perceived quality evaluation metrics. To achieve efficient AaaS and maximize the quality of generated content in wireless edge networks, we propose a deep reinforcement learning-enabled algorithm for optimal ASP selection. Simulation results show that the proposed algorithm can provide a higher quality of generated content to users and achieve fewer crashed tasks by comparing with four benchmarks, that is, overloading-avoidance, randomness, round-robin policies, and the upper-bound schemes.

Introduction

Artificial Intelligence-Generated Content (AIGC) techniques have gained significant attention due to the unprecedented ability to automate the creation of various content [1], for example, text, images, and videos. Undoubtedly, AIGC will significantly impact daily applications, especially Metaverse. With the ability to produce efficiently large amounts of high-quality content, AIGC can save time and resources that would otherwise be spent on manual content creation.

Recent research has made significant strides in the field of AIGC. Specifically, in text generation, the authors in [2] and [3] explored the application of deep learning techniques to generate coherent and diverse texts. For image generation, studies such as [4] and [5] focused on generating photo-realistic images using generative adversarial networks (GANs). In the audio generation, the authors in [6] explored deep learning techniques for synthesizing high-quality speech. OpenAI has been at the forefront of language model development, starting with the release of the GPT-3 model. This versatile model was capable of diverse tasks such as machine translation, text generation, and semantic analysis [7]. Recently, its successor, GPT-4, further pushed the envelope by demonstrating enhanced factual correctness and adherence to ethical guidelines. Concurrently, diffusion model-based DALL-E2 was introduced and recognized as a leading image generation model, surpassing GANs [8]. Supported by a dataset of 11 million images and over 1 billion masks, Meta Research's Segment Anything Model (https://segment-anything.com/) boasters impressive zero-shot inference capabilities, thereby shaping a new benchmark for instance segmentation tasks.

Despite these advancements, challenges persist in the field of AIGC. The models require extensive data for training and are difficult to deploy due to their size. For example, *Stability AI*, responsible for the Stable Diffusion model, maintains over 4,000 NVIDIA A100 GPU clusters and has spent over $50 million in operating costs (https://stability.ai/). A single training session of Stable Diffusion V1 requires 150,000 A100 GPU hours. Furthermore, the suitability of AIGC models varies based on the training datasets used. An AIGC model trained on a human face dataset, while effective at repairing corrupted face images, may not perform as well with blurred landscape images. Due to the diversity of users' tasks and the limited edge device capabilities, it is difficult to deploy multiple AIGC models on every network edge device. To tackle this challenge, we adopt the "Everything-as-a-service" (EaaS) concept and propose a new deployment scheme called "AIGC-as-a-service" (AaaS), which allows AI models to be deployed on edge servers and provides

Hongyang Du and Dusit Niyato are with Nanyang Technological University, Singapore; Zonghang Li is with the University of Electronic Sciences and Technology of China, China; Jiawen Kang (corresponding author) is with Guangdong University of Technology, China; Zehui Xiong is with Singapore University of Technology and Design, Singapore; Xuemin Shen is with the University of Waterloo, Canada; Dong In Kim is with Sungkyunkwan University, South Korea.

Digital Object Identifier: 10.1109/MWC.004.2300015

users with AIGC services. In the AaaS scheme, AIGC service providers (ASPs) deploy AI models on more powerful servers rather than on user-side mobile devices, offering instant services to users over wireless networks. This enables users to access and enjoy AIGC with low latency and resource consumption. There are several advantages of deploying AaaS in edge networks.

Personalization: AIGC models can generate content tailored to each user's requirements, providing a personalized and engaging experience. For example, personalized generative-AI product recommendations can be offered to users based on their locations, preferences, and usage patterns.

Efficiency: Deploying AIGC services closer to users improves the quality of service (QoS), for example, lower delay, and enables more efficient utilization of network and computing resources due to local content transfer.

Flexibility: AIGC can be customized and optimized to meet dynamic demands and resource availability. By scheduling wireless network users' access for ASPs, the overall QoS for users in the network can be maximized.

Thus, edge-enabled AaaS has the potential to revolutionize the way content is generated and delivered over wireless networks by addressing the deployment challenges. However, the current research on AIGC primarily focuses on AIGC model training, overlooking the resource allocation issues when deploying AIGC in wireless edge networks. AIGC models may require significant bandwidth and computing power to generate and deliver content to users, potentially leading to degraded network performance. Furthermore, scaling AaaS to accommodate a large number of users can be challenging. On the one hand, users aim to choose the ASPs with the best performance. On the other hand, it is important to avoid overloading certain AIGC services and requiring re-transmissions. *To the best of our knowledge, this is the first research work to discuss the deployments, aforementioned challenges, and future directions of AIGC in wireless edge networks.* Our contributions can be summarized as follows:

- We provide a comprehensive overview of the AIGC and techniques behind it. Then, we discuss various applications of AIGC and their use cases in wireless edge networks and their deployment challenges.
- We review the existing image-based perceived quality metrics. By conducting real experiments, we propose a general model to reveal the relationship between computational resource consumption and the quality of generated content in AaaS.
- We propose a deep reinforcement learning (DRL)-enabled method to achieve a dynamic selection of optimal ASPs. We demonstrate the superiority of our proposed DRL-enabled algorithm compared with four solutions, including upper-bound, overloading-avoidance, random, and round-robin policies.

AI-Generated Content and Techniques

In this section, we review the recent progress of AIGC. Specifically, we introduce the technologies behind the AIGC. Then, we discuss several categories of AIGC and associated applications in edge networks.

Generative-AI Techniques

We introduce generative-AI techniques in training AIGC models [9]. The basic model structures are shown on the left of Fig. 1.

Autoregressive Models (ARMs): ARMs belong to statistical modeling that involves predicting the future values of a time series based on past values [9]. ARMs can generate text or other media types for content generation by predicting the next element based on the previous ones. A potential use case for ARMs is to generate music by predicting the next note in a musical sequence based on the previous notes from edge users.

Variational Autoencoders (VAEs): VAEs can generate new data by learning a compact, latent representation of the input data, consisting of an encoder network and a decoder network [9]. The encoder network processes the input data and outputs a latent representation. The decoder network takes this latent representation as input and generates synthetic data similar to the input data.

Generative Adversarial Networks (GANs): GANs consist of two neural networks, that is, generator and discriminator networks [4]. The two networks are trained together to improve the generator's ability to generate realistic images and the discriminator's ability to distinguish synthetic images from real images.

Flow-Based Models (FBMs): FBMs transform a simple distribution into a target distribution through a series of invertible transformations [9]. These transformations are implemented as neural networks, and the process of applying the transformations is referred to as "flow."

Diffusion Models (DMs): DMs are trained to denoise images blurred by Gaussian noise to learn how to reverse the diffusion process [8]. Several diffusion-based generative models have been proposed, including diffusion probabilistic models, noise-conditioned score networks, and denoising diffusion probabilistic models.

Moreover, classic techniques such as Transformer can also be used in AIGC models, which are discussed in the following.

Categories of AIGC and Applications in Mobile Networks

We then present several categories of AIGC technologies and their applications in edge networks, which can serve as potential future research directions.

Text-to-Text AIGC: Text-to-text AIGC can generate the human-like message as an output based on a given text input. Therefore, it can be used for automatic answers, language translation, or article summarization. One representative text-to-text AIGC model is the Chat Generative Pre-training Transformer (GPT) (https://openai.com/blog/chatgpt/), a language model developed by OpenAI [7]. The GPT is trained on a large dataset of human-generated text, such as books or articles. The model can then create text by predicting the next word in a sequence based on the words that come before it. GPT has been highly successful and has achieved state-of-the-art results on several natural language processing (NLP) benchmarks. In wireless edge networks, as shown in Fig. 1, GPT can serve as a chatbot that provides drivers with navigation and information alert services. Additionally, real-time services like instant language

AIGC models can generate content tailored to each user's requirements, providing a personalized and engaging experience.

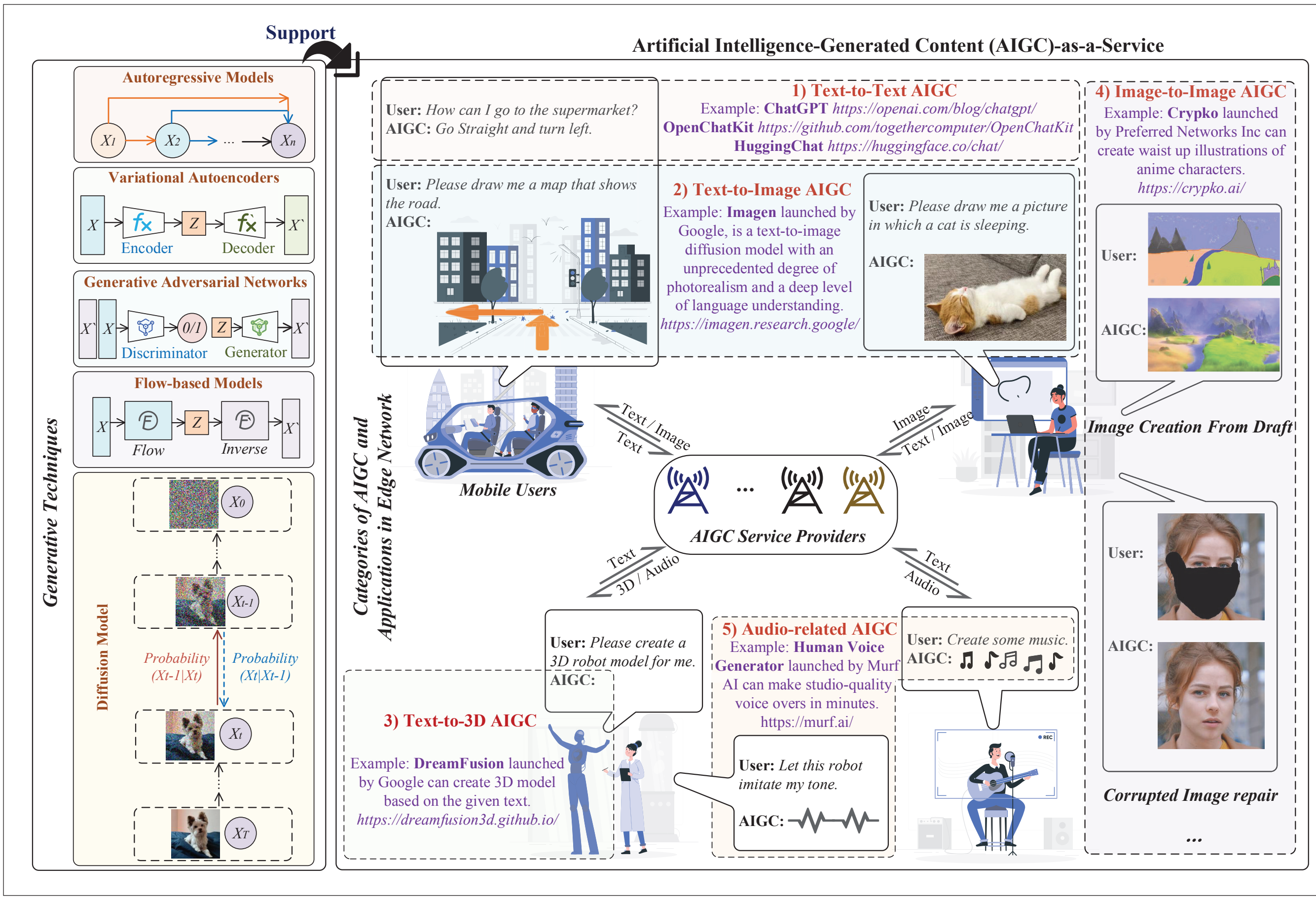


FIGURE 1. Generative techniques in AIGC [9], categories of AIGC, and applications in wireless edge networks. The listed AIGC services, for example, ChatGPT for text-to-text AIGC (https://openai.com/blog/chatgpt/), Imagen for text-to-image AIGC (https://imagen.research.google/), DreamFusion for text-to-3D AIGC (https://dreamfusion3d.github.io/), Crypko for image-to-image AIGC (https://crypko.ai/), and Human Voice Generator for audio-related AIGC (https://murf.ai/), are provided as examples of existing applications, illustrating the potential use cases for our proposed AIGC-as-a-service (AaaS) deployment in wireless edge networks.

translation, for example, Translation AI (https://cloud.google.com/translate), can be deployed. For instance, during live video conferences, real-time language translation services are crucial for seamless communication.

Text-to-Image AIGC: Text-to-image AIGC allows users to generate images based on text input, enabling the creation of visual content from written descriptions. It can be regarded as a combination of natural language processing and computer vision techniques. As shown in Fig. 1, the text-to-image AIGC can assist mobile users with various activities. For example, users in Internet-of-Vehicles can request visual-based path planning. Furthermore, text-to-image AIGC can also assist users in creating art and producing pictures in various styles based on users' descriptions or keywords.

Text-to-3D AIGC: Text-to-3D AIGC can generate 3D models from text descriptions while using wireless AR applications. Typically, generating 3D models requires higher computational resources than generating 2D images. Considering the development of next-generation Internet services, such as Metaverse [10], generating 3D models based on text without complicated manual design is a fascinating prospect. Considering the development of next-generation Internet services, such as Metaverse [10], generating 3D models based on text without complicated manual design is a fascinating prospect. Real-time AR experiences, integrated with text-to-3D AIGC, allow users to interact with and visualize 3D content instantly. Empowering these services by edge nodes is crucial for meeting AR applications' low-latency and real-time processing demands, ensuring a seamless and immersive experience while highlighting the benefits of AIGC services in wireless edge networks.

Image-to-Image AIGC: Image-to-Image AIGC uses AI models to generate realistic images from source images or create a stylized version of an input image. For example, when it comes to assisting artwork creation, image-to-image AIGC can generate visually satisfying pictures based solely on user-inputted sketches. Furthermore, image-to-image AIGC can be used for image editing services. Users can remove occlusions in one image or repair corrupted images.

Audio-related AIGC: Audio-related AIGC models analyze, classify, and manipulate audio signals, including speech and music. Specifically, text-to-speech models are designed to synthesize natural-sounding speech from text input. Music generation models can synthesize music in a variety of styles and genres. Audio-visual music generation involves using both audio and visual information, such as music videos or album artwork, to generate music compositions that are more closely tied to a particular visual style or theme. Moreover, audio-related AIGC can serve as voice assistants that answer users' queries. Alexa (https://developer.amazon.com/en-US/alexa) and Siri (https://www.apple.com/sg/siri/) are examples of real-life applications.

Given the power of AIGC models, there are several challenges in deploying AaaS in wireless edge networks, which are introduced in the following.

AI-Generated Content-as-a-Service in Wireless Edge Networks

In this section, we discuss the AaaS in detail, including the challenges and performance metrics.

AI-Generated Content-as-a-Service and Challenges

To deploy AaaS in wireless edge networks, the ASPs should first train AIGC models on large datasets. The AIGC models would need to be hosted on edge servers and can be accessed by users. Continuous maintenance and updates would be required to ensure that the AIGC models remain accurate and effective for generating high-quality content. Users can submit requests for content generation and receive the generated content from edge servers rented by ASPs. Despite several advantages of deploying AaaS in wireless edge networks, there are pertaining challenges to be addressed.

Bandwidth Consumption: The AIGC consumes a significant amount of bandwidth. Particularly for high-resolution images, both upload and download processes require considerable network resources. For example, the data size of an AI-generated wallpaper in wallhaven (https://wallhaven.cc/tag/133451) can be around 10 Megabytes. Furthermore, deploying AIGC models at the edge requires resource management to avoid network congestion and ensure low-latency services.

Time-Varying Channel Quality: The QoS in AaaS can be affected by the wireless transmission of the generated content. Low Signal-to-Noise Ratio (SNR), high Outage Probability (OP), and high bit-error probability (BEP) can degrade QoS of AIGC services and users' satisfaction, which is due to time-varying fading channels when the channel encounters deep fading occasionally.

Computation Resource Consumption: The deployment of AIGC models at the network edge requires efficient utilization of limited computational resources on edge devices, for example, fine-tuning and inference. For example, the quality of the output of the diffusion model-AaaS increases with the number of inference steps.

Utility Maximization and Incentive Mechanism: Incentive mechanism design is important in AaaS as it can motivate ASPs to generate high-quality content, meeting the target goals and objectives. Here, the utility function should represent the perceived QoS from users.

Edge-Aware Content Generation: The dataset used for training AIGC models can impact the quality of the generated content. To maximize user satisfaction, edge-enabled AIGC services should be designed to consider the specific requirements of users in edge environments, such as context-awareness and location-based personalization.

A prevalent challenge in addressing the issues associated with AaaS is evaluating their performance. Although numerous evaluation metrics have been proposed across various modalities, most are either AI model-based or difficult to compute, lacking a straightforward mathematical expression. For the optimal design of AaaS in wireless networks, AI-enabled resource allocation solutions can leverage performance values derived from AI models to account for users' subjective experiences. However, traditional mathematical resource allocation schemes necessitate modeling the relationship between computational resource consumption, such as the number of inference steps in the diffusion model, and the quality of the generated content, as depicted in Fig. 2. To tackle this problem, we use image-related AaaS as an example, introducing various performance evaluation metrics and investigating the mathematical relationship between metric values and computational resource consumption in subsequent sections.

Performance Metric Modelling

We first discuss AIGC evaluation metrics, focusing on the assessment of perceived image quality. We then formulate the relationship between computational resource consumption and the quality of generated content in AaaS.

Image-Based Metrics: The image quality assessment metrics can be distribution-based and image-based. The distribution-based metrics, for example, Frechet inception distance [11], take a list of image features to compute the distance between distributions for evaluating generated images. However, for practical AaaS, the quality evaluation is subjective, and it is hard for users to calculate distribution-based metrics. Thus, we focus on image-based metrics that attempt to achieve consistency in quality prediction by modeling salient physiological and psycho-visual features of the human visual system or by signal fidelity measures. Specifically, without access to the original image as a reference, *no-reference image quality evaluation* methods can be considered [11].

Total Variation (TV): TV is a measure of the smoothness of an image. One common way to compute total variation is to take the sum of the absolute differences between adjacent samples in an image. This measures the "roughness" or "discontinuity" of the image.

Blind/Referenceless Image Spatial Quality Evaluator — BRISQUE: (http://live.ece.utexas.edu/research/quality/) BRISQUE utilizes scene statistics of locally normalized luminance coefficients to quantify possible losses of "naturalness" in the image due to distortions [12]. It has been shown that BRISQUE performs well in correlation with human perception of quality.

The higher the image quality, the smaller the values of TV and BRISQUE.

For AaaS where a reference image is available, we can use *full-reference image quality evaluation* methods [11].

Discrete Cosine Transform Subbands Similarity (DSS): DSS exploits essential characteristics of human visual perception by measuring changes in structural information in subbands in the discrete cosine transform (DCT) domain and weighting the quality estimates for these subbands [13].

Haar Wavelet-based Perceptual Similarity Index (HaarPSI): HaarPSI utilizes the coefficients obtained from a Haar wavelet decomposition to assess local similarities between two images, as well as the relative importance of image areas.

The AIGC consumes a significant amount of bandwidth. Particularly for high-resolution images, both upload and download processes require considerable network resources.

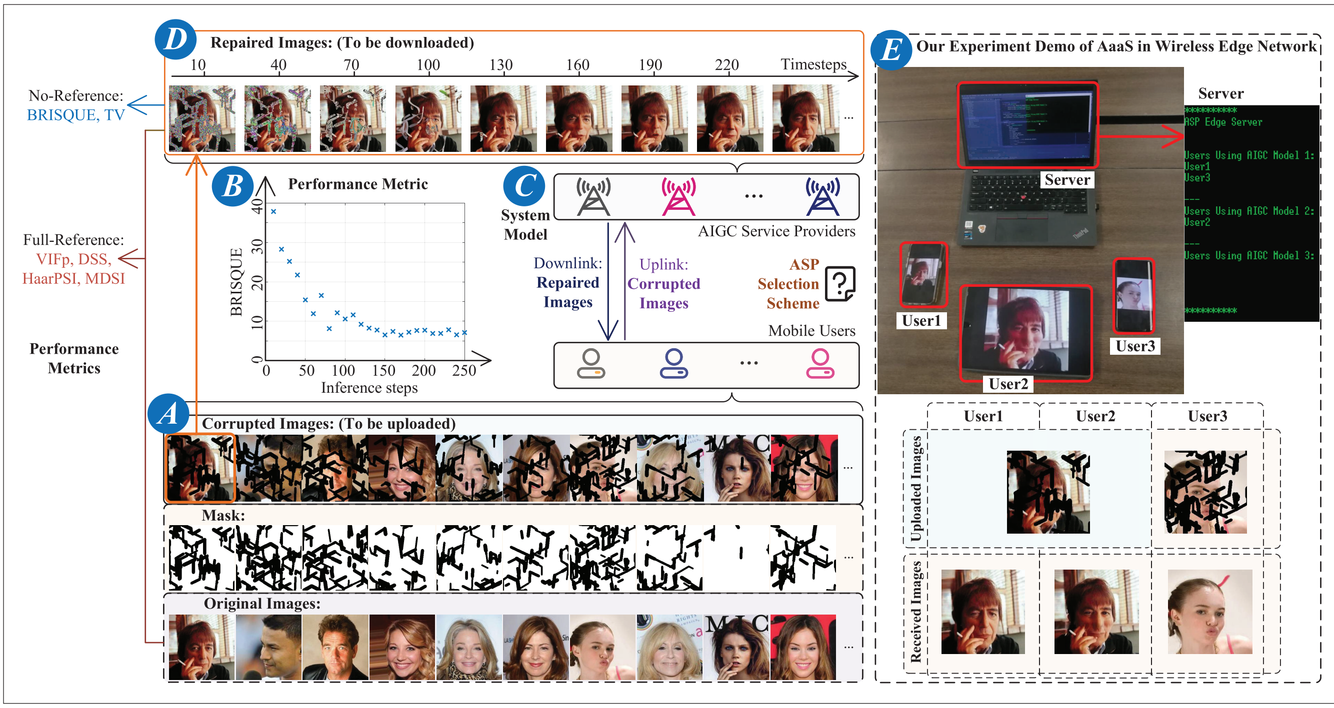


FIGURE 2. Example of an AaaS for repairing corrupted images. The corrupted images are shown in Part A, and the repaired images under different inference steps are shown in Part D. Part B shows how the performance metric, BRISQUE, varies over different inference steps. Part C shows the system model of ASP selection problem. A experiment demo is shown in Part E.

Mean Deviation Similarity Index (MDSI): MDSI is a reliable and complete reference perceptual image quality assessment model that utilizes gradient similarity, chromaticity similarity, and deviation pooling.

Visual Information Fidelity (VIF): VIF is a competitive way of measuring fidelity that relates well with visual quality, which quantifies the information in the reference image and how much of the reference information can be extracted from the distorted image.

The higher the image quality, the higher the values of the aforementioned full-reference image quality metrics.

A General Modelling of Perceived Image Quality Metric: AIGC models based on diffusion models are becoming mainstream. As shown in the left of Fig. 1, the diffusion process can be regarded as a step-wise denoising process. Thus, increasing the number of inference steps will improve the perceived image quality. However, the generated image quality does not always increase with the number of steps. Excessive inference steps incur unnecessary resource consumption. We conduct real experiments to investigate the relationship between the number of inference steps and various perceived image quality metrics, that is, TV, BRISQUE, DSS, HaarPSI, MDSI, and VIF.

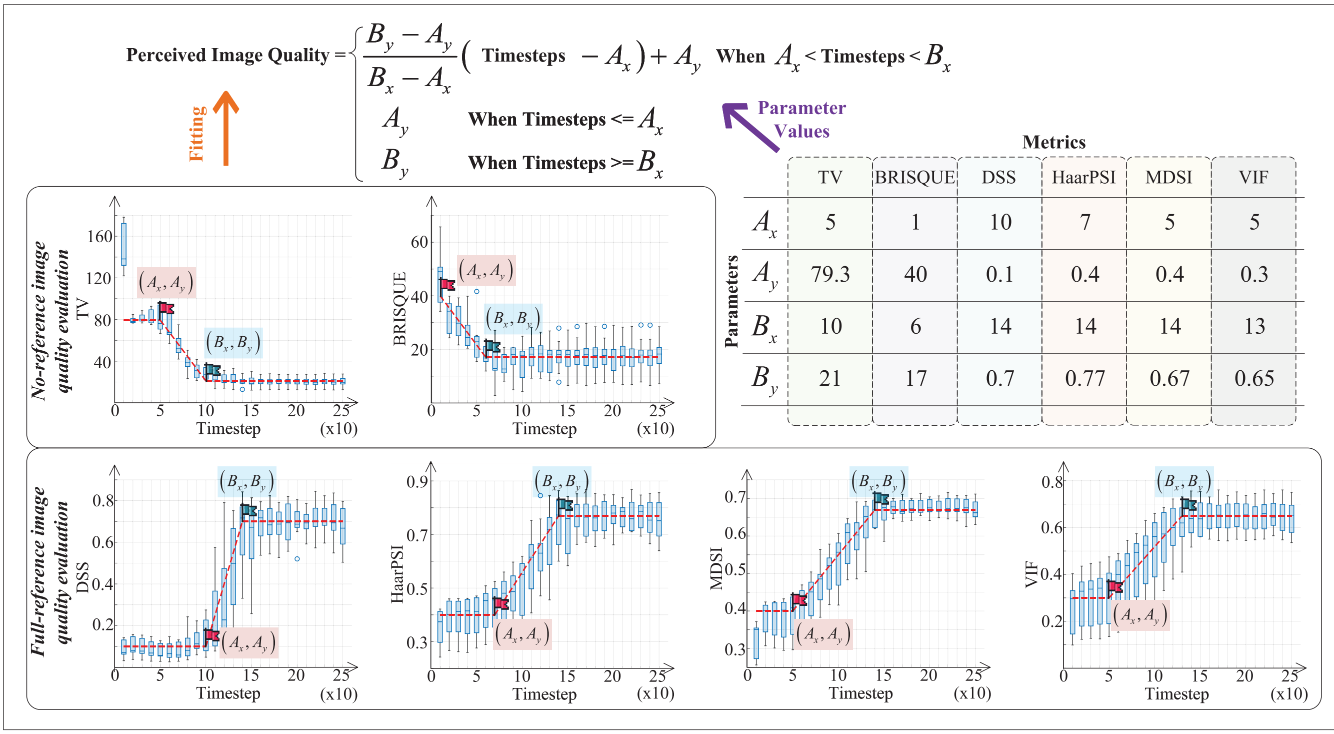
The experimental platform is built on an Ubuntu 20.04 system with an AMD Ryzen Threadripper PRO 3975WX 32-Cores CPU and an NVIDIA RTX A5000 GPU. We take diffusion model-based corrupted image restoration service as an example of AaaS. Specifically, we deploy the well-trained model, *RePaint*, proposed in [14] on our server. As shown in Fig. 2 (Part A), we first generate a series of corrupted images, for example, 20 images, with the help of masks. Then, these corrupted images are fed into *RePaint*. We can observe that the corrupted image gradually recovers as the inference progresses, as shown in Fig. 2 (Part D). Moreover, the values of image quality metric, for example, BRISQUE, decrease, as shown in Fig. 2 (Part B). We show the values of each performance metric under the different number of timesteps in Fig. 3.

Thus, we present a general model of the perceived image quality metric that contains four parameters, as shown at the top of Fig. 3. Specifically, A_x is the minimum number of inference steps where the image quality starts to improve, A_y is the lower bound of the image quality, which can be regarded as the evaluation value for images with high noise, B_x is the number of inference steps when the image quality starts to stabilise because of the capability of AIGC models, and B_y is the highest image quality value that the model can achieve. Regardless of whether the performance metric value is positively or inversely proportional to the image quality, and regardless of the AaaS types, we can easily find the points (A_x, A_y) and (B_x, B_y) experimentally, as shown in Fig. 3.

Lesson Learned: *Despite the inherent uncertainty of the diffusion process, from Fig. 3, we can observe that the perceived image quality increases or decreases approximately proportionally with the increase of inference steps. In the practical AIGC model analysis, we can perform experiments with the simple fitting method as shown in Fig. 3 to a performance metric to obtain four parameters in our proposed general mathematical model. Then, the model can be used in wireless edge network-enabled AIGC services analysis.*

Deep Reinforcement Learning-Aided Dynamic ASP Selection

In this section, we study the optimal ASP edge server selection problem. We propose a DRL-en-



	TV	BRISQUE	DSS	HaarPSI	MDSI	VIF
A_x	5	1	10	7	5	5
A_y	79.3	40	0.1	0.4	0.4	0.3
B_x	10	6	14	14	14	13
B_y	21	17	0.7	0.77	0.67	0.65

FIGURE 3. The relationship between the number of inference steps and various perceived image quality metrics, that is, TV, BRISQUE, DSS, HaarPSI, MDSI, and VIF. A general model of the perceived image quality metric that contains four parameters is presented.

abled solution to maximize utility function while satisfying users' requirements.

AaaS System Model

Our demo is shown in Fig. 2 (Part E). Specifically, three users are selecting between two image reparation AIGC models that are trained on datasets CelebA-HQ and Places2 [14], respectively. User 1 and User 2 upload the same corrupted images. We can observe that different AIGC models will create different results for the same user task.

In this study, we examine the case for large-scale deployment of AaaS in wireless edge networks through a simulation involving 20 ASPs and 1,000 edge users. Each ASP offers AaaS with a maximum resource capacity, defined as the total number of diffusion timesteps within a time window, which ranges from 600 to 1,500 at random. Users submit multiple AIGC task requests to the ASPs at different times, specifying the amount of AIGC resources they require, that is, diffusion timesteps, randomly set between 100 and 250.[1] The quality of AIGC models provided by different ASPs varies, as assessed by the BRISQUE metric. User task arrivals follow a Poisson distribution, with a rate of λ = 0.288 hour/request over a period of 288 hours, resulting in a total of 1,000 tasks. Note that quality of AIGC models of different APSs varies, as some repaired images may appear more realistic and natural.

A straightforward yet less effective ASP selection strategy involves users sending task requests directly to the ASP that can offer the highest quality of generated content. However, this approach can overload some ASPs due to insufficient computational resources, resulting in interrupted tasks in practice. Furthermore, the quality of generated content from ASPs is generally unknown to users. Mobile users would need to query ASPs multiple times to estimate content quality for executing the myopic selection, which introduces unnecessary load and consumes wireless network resources. Therefore, identifying a suitable ASP for user tasks to maximize the overall system utility while minimizing AIGC resource overload and interruptions caused by popular ASPs when content quality is unknown is a challenging and critical problem to address.

Deep Reinforcement Learning-Based Solution

We use the soft actor–critic (SAC) DRL [15] to solve the above dynamic ASP selection problem. As shown in Fig. 4, the learning process alternates between policy evaluation (Critic) and policy improvement (Actor). Unlike the conventional actor-critic architecture, the policy in SAC is trained to maximize a trade-off between the expected return and entropy. The state space, action space, and reward in the AaaS environment are defined as in the following.

State: The state space is composed of two parts:
- A feature vector (*the demand of AIGC resources for current user task and the estimated completion time of the task that is calculated based on the diffusion timestep requirement*) of the arriving user task
- Feature vectors (*the total AIGC resources of the i-th ASP and the currently available resources of the i-th ASP*) of all ASPs in the current state.

Action: The action space is an integer indicating the selected ASP. In detail, the actor policy network outputs a 20-dimensional logits vector, and then the probability of selecting each ASP is obtained after being post-processed by the softmax operator. Finally, DRL selects an ASP to handle the current user task according to the assigned probability of each ASP.

Reward: The reward consists of two parts: a quality of generated content reward and a con-

[1] We consider that users can specify the number of diffusion timesteps they require, similar to the functionality offered in the Stable Diffusion v1-5 Demo by Hugging Face, available at https://huggingface.co/spaces/runwayml/stable-diffusion-v1-5.

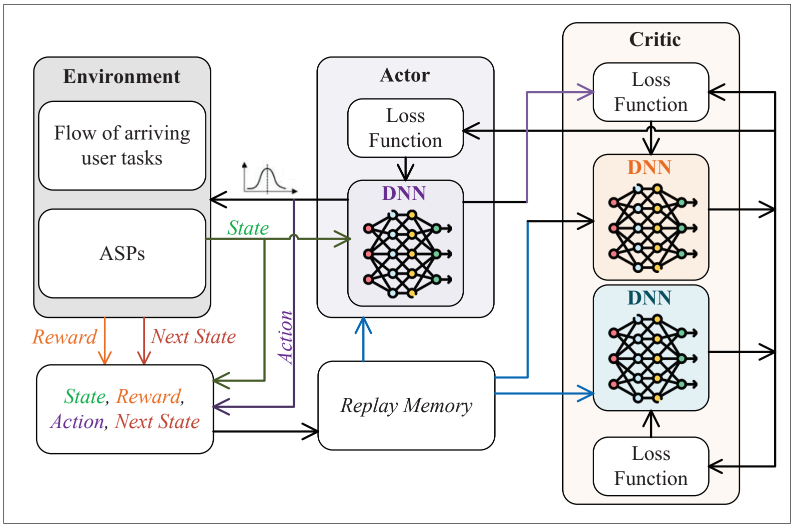


FIGURE 4. The structure of soft actor-critic DRL algorithm. The actor network is used for action selection and the critic networks are used for reward estimation.

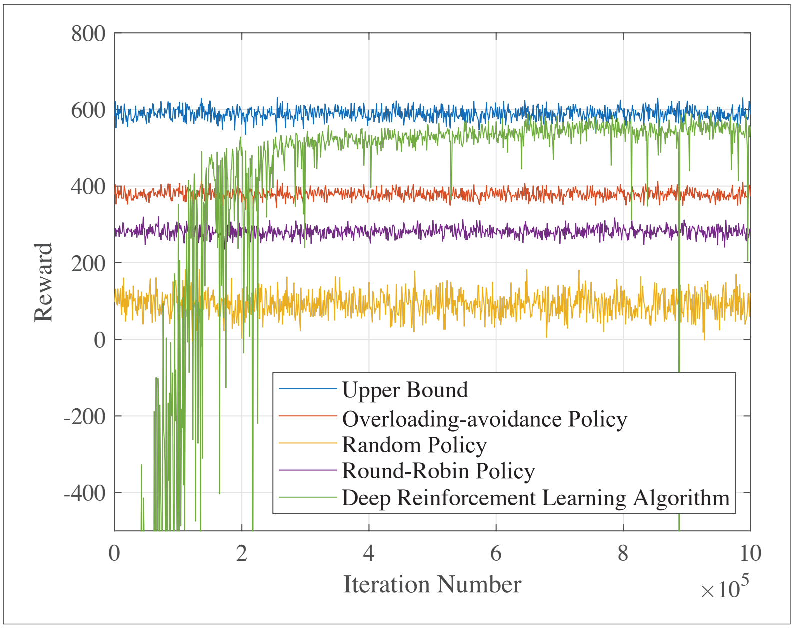


FIGURE 5. Reward values versus the number of iteration number in DRL. For performance comparison, we show the results of overloading-avoidance, random, round-robin policies, and their upper bound.

gestion penalty. The former is defined as the perceived quality of the repaired image, as discussed earlier. Furthermore, any action that overloads AIGC models must be penalized as a congestion penalty. First, the action itself should be punished with fixed *penalty value*. Second, considering that ill-considered actions can cause bottleneck ASP's model to crash and the running tasks will be interrupted, the current action will also be subject to additional penalties according to the progress of ongoing tasks. The total reward returned is the quality reward minus the congestion penalty. Note that a larger penalty value will encourage DRL to pay more attention to avoid crashes.

We compare the performance of the DRL-enabled ASP selection algorithm with four benchmarks. The lower bound is the random allocation policy, assigning every new user task to an ASP randomly. The optimal policy provides an approximation of the upper-bound performance. It operates under the assumption that quality scores for each task on all ASPs are known, which is rarely the case in practice due to the requirement for a posteriori knowledge. In this policy, we employ a greedy algorithm to allocate new user tasks, which selects the ASP with sufficient AIGC resources and the highest quality score for each new task. Note that it is not a brute-force method, as the greedy algorithm does not explore all possible combinations of task assignments. Furthermore, we implement the round-robin and overloading-avoidance policies, which are widely adopted in web applications to realize load balancing. It is simple, easy to implement, and starvation-free. The overloading-avoidance policy assigns the new user task to the ASP with most AIGC resources currently available to prevent or reduce the severity of overloads and crashes.

Figure 5 shows the utility curves (i.e., reward) of the DRL-enabled ASP selection policy and four benchmark policies. Since DRL can learn and evolve, as the learning step progresses, DRL has a more comprehensive and accurate selection of the ASP. Thus the utility rises rapidly, showing learning ability. One interesting finding is that when DRL overtakes the round-robin, DRL already has a specific load-balancing capability. Immediately afterward, DRL surpasses overloading-avoidance. At this time, DRL has learned to avoid actions that may cause crashes, thereby avoiding the congestion penalty. Then, DRL starts to learn the priority of different ASPs, and it seeks to place the current user task on the ASP with high quality to maximize the reward. Therefore, as shown in Fig. 5, DRL still has much room for improvement after surpassing the overloading-avoidance policy and finally reaching an episodic reward comparable to the upper-bound policy.

Figure 6 counts the episodic rewards, the average rewards of finished tasks, and the number of crashed tasks of the five policies. On the one hand, the DRL-enabled ASP selection policy achieves zero task crashes and minimizes the congestion penalty, which is critical to providing a satisfying quality of generated content to users. On the other hand, DRL policy can learn the quality of content that ASPs may provide, which is unknown in other policies. Then, DRL can assign user tasks to ASPs that can provide higher QoS so that the average reward for each task is effectively increased. Moreover, it is essential to address the energy implications of task crashes at the edge. The task crashes require user request re-submissions, leading to increased computation and energy consumption. This additional load impacts not only the user's device but also places strain on the edge network. By mitigating task crashes, the DRL-enabled ASP selection policy effectively reduces energy consumption resulting from re-submissions and improves the system's overall energy efficiency.

Future Direction

Secure AIGC-as-a-Service

When deploying AaaS in a wireless network, the requests from users and the generated contents from ASPs are transmitted in a wireless environment. Therefore, advanced security techniques for AIGC need to be studied, for example, protecting the transmission of AIGC data through improved physical layer security techniques. Moreover, blockchain can be used to enable decentralized content distribution, allowing content to be shared and accessed directly between users without the need for a central authority. The authenticity and provenance of AIGC can be verified with the aid of blockchain, helping to ensure that the AIGC is accurate and trustworthy. Furthermore, during the training process of AIGC models, the privacy of the training data needs to be guaranteed, especially biometric data, such as face images. One possible solution is to train the model by secure federated learning.

IoT-Based and Wireless Sensing-Aided Passive AaaS

Considering the fast development of sensing technologies, we aim to enable ubiquitous passive AaaS with wireless sensing signals. For example, wireless sensors can gather data about the environment or user behavior, which can then be fed into an AIGC model to generate relevant content. Wireless sensing-aided passive AaaS can also be used in healthcare. Specifically, by using IoT devices to detect users' activity levels, sleep patterns, or heart rates with the aid of wireless sensing, the AIGC could generate content such as personalized workout plans. Moreover, the mobility of network devices predominantly affects the throughput of the connected links for AaaS, which is worth further study.

Personalized Resource Allocation in AaaS

Although the current AIGC models can meet users' needs with customized tasks, more studies are needed to achieve personalized AIGC services. For example, for text-to-image AaaS, when both users enter the text "A monkey is standing next to a zebra," current ASPs produce similar images for users. However, if we deduce that the two users are a horse trainer and a monkey researcher, respectively, we can personalize the computing resource allocation [10]. Specifically, more resources can be allocated to generate and transmit the zebra in the image for the horse trainer. For the monkey researcher, the AIGC model that is more adapted to generating monkey images should be assigned to handle the task. One potential solution is incorporating user feedback and preferences into the content generation process and developing techniques for evaluating the effectiveness of personalized content.

Conclusion

We reviewed the AIGC technologies and discussed the applications in wireless networks. To provide AIGC services to users, we proposed the concept of AaaS. Then, the challenges of deploying AaaS in wireless networks are discussed. In addressing these challenges, a fundamental problem is about the mathematical relationship between the resource consumption and the perceived quality of the generated content. After exploring various image-based performance evaluation metrics, we proposed a general modeling equation. Moreover, we studied the important ASP selection problem. A DRL-enabled algorithm was used to achieve nearly optimal ASP selection. As the first article to discuss AIGC in wireless networks, we hope that this article can inspire researchers to contribute to the advancement of wireless edge networks-enabled AaaS.

Acknowledgment

This research is supported by NSFC under grant No. 62102099, U22A2054, and the Pearl River Talent Recruitment Program under Grant 2021QN02S643, and Guangzhou Basic Research Program under Grant 2023A04J1699.

This research is also 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), 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 research is also supported by the SUTD SRG-ISTD-2021-165, the SUTD-ZJU IDEA Grant (SUTD-ZJU (VP) 202102), and the Ministry of Education, Singapore, under its SUTD Kickstarter Initiative (SKI 20210204).

This research was supported in part by the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience program (IITP-2020-0-01821) and the ITRC (Information Technology Research Center) support program (IITP-2023-RS-2023-00258639) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

H. Du and Z. Li have equal contributions.

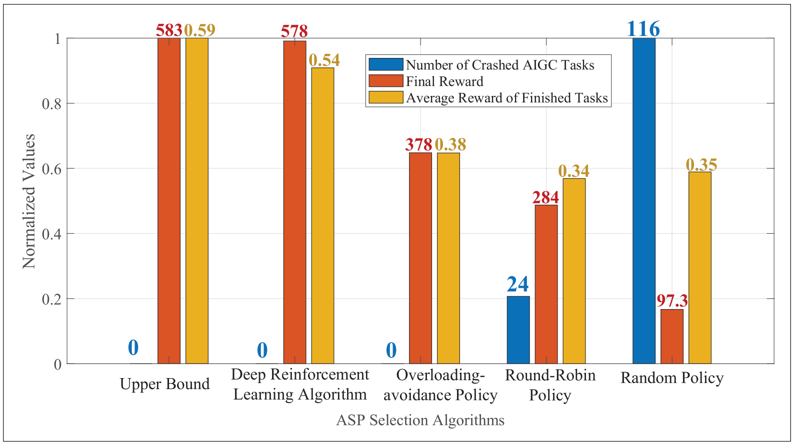


FIGURE 6. Comparison of episodic rewards, average rewards of finished tasks, and number of crashed tasks.

References

[1] L. Yunjiu, W. Wei, and Y. Zheng, "Artificial Intelligence-Generated and Human Expert-Designed Vocabulary Tests: A Comparative Study," *SAGE Open*, vol. 12, no. 1, Jan. 2022.
[2] M. Chen *et al.*, "Generative Pretraining From Pixels," *Proc. Int'l. Conf. Mach. Learn. PMLR*, 2020, pp. 1691–1703.
[3] J. Guo *et al.*, "Long Text Generation via Adversarial Training With Leaked Information," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.
[4] T. Karras *et al.*, "Progressive Growing of Gans for Improved Quality, Stability, and Variation," *Proc. Int'l. Conf. Mach. Learn.*, 2018.
[5] X. Huang *et al.*, "Multimodal Unsupervised Image-to-Image Translation," *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–89.
[6] W. Ping *et al.*, "WaveFlow: A Compact Flowbased Model for Raw Audio," *Proc. Int'l. Conf. Mach. Learn. PMLR*, 2020, pp. 7706–16.
[7] L. Floridi and M. Chiriatti, "GPT-3: Its Nature, Scope, Limits, and Consequences," *Minds Mach.*, vol. 30, no. 4, Apr. 2020, pp. 681–94.
[8] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," *Adv. Neural Inf. Process. Syst.*, vol. 34,

2021, pp. 8780–94.
[9] G. Harshvardhan *et al.*, "A Comprehensive Survey and Analysis of Generative Models in Machine Learning," *Comput. Sci. Rev.*, vol. 38, 2020, p. 100285.
[10] H. Du *et al.*, "Attention-Aware Resource Allocation and QoE Analysis for Metaverse xURLLC Services," *IEEE JSAC*, vol. 41, no. 7, 2023.
[11] S. Kastryulin, D. Zakirov, and D. Prokopenko, "PyTorch Image Quality: Metrics and Measure for Image Quality Assessment," 2019; opensource software available at https://github.com/photosynthesis-team/piq.
[12] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, Dec. 2012, pp. 4695–4708.
[13] L. Gatys, A. Ecker, and M. Bethge, "A Neural Algorithm of Artistic Style," *J. Vis.*, vol. 16, no. 12, Dec. 2016, pp. 326–26.
[14] A. Lugmayr *et al.*, "Repaint: Inpainting Using Denoising Diffusion Probabilistic Models," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11,461–71.
[15] P. Christodoulou, "Soft Actor-Critic for Discrete Action Settings," arXiv preprint arXiv:1910.07207, 2019.

Biographies

Hongyang Du 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.

Zonghang Li is a Ph.D. candidate at the School of Information and Communication Engineering at the University of Electronic Science and Technology of China (UESTC). His research interests include intelligent communication and computing systems, distributed machine learning, and federated learning. He was awarded the 2021 Leading and Innovative Technology Achievement Award by the China Institute of Communications (CIC). He was a visiting scholar at Nanyang Technological University and Oxford University.

Dusit Niyato 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.

Jiawen Kang 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.

Zehui Xiong is an Assistant Professor at Singapore University of Technology and Design, and also an Honorary Adjunct Senior Research Scientist with Alibaba-NTU Singapore Joint Research Institute, Singapore. He received the PhD degree in Nanyang Technological University (NTU), Singapore. His research interests include wireless communications, blockchain, edge intelligence, and Metaverse.

Xuemin (Sherman) Shen received the PhD degree in electrical engineering from Rutgers university, New Brunswick, New Jersey, in 1990. He is currently 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.

Dong In Kim received the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1990. He is a Professor with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea. His research interests include internet of things, wireless power transfer, and connected intelligence.