QuAsyncFL: Asynchronous Federated Learning with Quantization for Cloud-Edge-Terminal Collaboration Enabled AIoT

Ye Liu, *Member, IEEE* Peishan Huang, Fan Yang, Kai Huang, and Lei Shu *Senior Member, IEEE*

***Abstract*—Federated Learning is a promising technique that facilitates cloud-edge-terminal collaboration in Artificial Intelligence of Things (AIoT). It will enable model training without centralizing data, addressing privacy and security concerns. However, when applied to AIoT, this technique faces several challenges, such as low communication efficiency among terminal devices, edges, and cloud platforms. In this paper, we propose a novel approach called QuAsyncFL, which combines asynchronous federated learning with an unbiased nonuniform quantizer to address the issue of low communication efficiency. Moreover, we provide a detailed theoretical analysis of convergence with quantized gradients proving that the model could converge to a certain bound. Our experiments demonstrate that QuAsyncFL outperforms the original approach, achieving significant improvements in terms of communication efficiency. The research results represent a further step towards developing cloud-edge-terminal collaboration enabled AIoT.**

***Index Terms*—Artificial Intelligence of Things, Cloud-Edge-Terminal Collaboration, Asynchronous Federated Learning, Quantization, Communication Efficiency**

I. Introduction

Artificial Intelligence of Things (AIoT) [1] is an emerging technology that has recently captured considerable attention. It leverages the capabilities of Artificial Intelligence (AI) and the IoT to develop more intelligent and efficient network systems that can automate tasks and improve decision-making processes. In AIoT systems, AI algorithms are used to analyze and interpret the massive amount of data generated by IoT devices, resulting in more accurate predictions, faster response times, and improved system performance [2]. AIoT has numerous potential applications, such as smart cities [3] and smart agriculture [4], [5]. Therefore, it is anticipated to be a promising paradigm that has the potential to solve complex problems in various domains and revolutionize our society.

While AIoT is promising, the development is still in its early stages, and there are several challenges that need to be addressed for reaching its full potential through cloud-edge-terminal collaboration [6]. One of the primary challenges is data privacy and security. With the increasing volume and variety of data exchanging among cloud, edge, and IoT devices, ensuring the privacy and security of this data is crucial for the success of AIoT. The use of AI algorithms to analyze and interpret this data further highlights the need for secure and private data management. In addition, as more devices are connected to AIoT in cloud-edge-terminal collaboration architecture, efficient communication of the huge traffic is essential for the widespread adoption of AIoT. Moreover, the lack of interoperability and transparency between heterogeneous networks is another challenge that limits the integration and sharing of data in cloud-edge-terminal collaboration enabled AIoT, ultimately reducing its benefits.

Federated Learning [7] is expected to be a key enabler for cloud-edge-terminal collaboration enabled AIoT. This approach leverages cloud platforms to provide the infrastructure for storing and processing large volumes of data, while edge computing enables processing closer to the source of the data. Terminal IoT devices deliver the end-user interface and contribute to the training of AI models. In this way, federated learning allows for the model training of AIoT without centralizing data. Instead, data remains on IoT devices, and the models are trained locally on them. Only the model updates are shared with the edge and cloud. This approach provides many benefits [8], including increased privacy [9], scalability, and efficiency, to address the abovementioned challenges.

To date, numerous studies have been conducted on federated learning, which can be categorized based on the distribution of data into three types: horizontal federated learning, vertical federated learning, and federated transfer learning. Horizontal federated learning [10] is suitable for datasets with a large overlap in features but few samples, while vertical federated learning [11] is applicable for datasets with many samples that share the same data identity but differ in feature space. Federated transfer learning [12] is useful when datasets differ not only in samples but also in feature space. On the other hand, federated learning can be categorized based on timeliness into two types: synchronous federated learning and asynchronous federated learning. Synchronous federated learning [13] means

The work is supported in part by the open Foundation of the Key Laboratory of Architectural Acoustic Environment of Anhui Higher Education Institutes under Grant No.AAE2021YB01, the Macau Young Scholars Program under Grant No.AM2021016, and National Natural Science Foundation of China under Grant 62102170. *(Corresponding author: Peishang Huang)*

Ye Liu is with Key Laboratory of Architectural Acoustic Environment of Anhui Higher Education Institutes, Hefei 230601, China, and College of Artificial Intelligence, Nanjing Agricultural University, Nanjing 210031, China (e-mail: yeliu@njau.edu.cn).

Peishan Huang is with the School of Computer Science and Engineering, Macau University of Science and Technology, Macau, China (e-mail:3220006969@student.must.edu.mo)

Fan Yang is with School of Mathematics and Statistics, Jiangsu Normal University, Xuzhou, 221116, China. (email: yangfan@jsnu.edu.cn).

Kai Huang is with Institute of Agricultural Facilities and Equipment, Jiangsu Academy of Agricultural Sciences, Nanjing, China (e-mail: kai_huang@njau.edu.cn).

Lei Shu is with Nanjing Agricultural University, Nanjing 210031, China, and with the School of Engineering, College of Science, University of Lincoln, Lincoln LN6 7TS, U.K. (e-mail: lei.shu@ieee.org).

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3290818

that participating devices synchronize with each other to update models and is suitable for low network latency and high bandwidth. Asynchronous federated learning [14] allows participating devices to update models independently and asynchronously, making it beneficial for high network latencies and bandwidth constraints. Asynchronous federated learning can also be advantageous for large-scale networks where synchronization of all devices can cause slower training times and network congestion.

In this paper, our focus is on asynchronous federated learning, which is more suitable for cloud-edge-terminal collaboration enabled AIoT. While it offers many advantages over synchronous federated learning, there are still many open issues to be addressed. One of its primary limitations is slower convergence of the model due to the potential for stale gradients. IoT devices may update the model with outdated information, leading to reduced accuracy and slower convergence. The combination of QuAsyncFL and cloud-edge collaboration can not only effectively improve convergence rate and communication efficiency, but also enable every device and its data in the IoT to participate in the learning of the information update. Therefore, the purpose of this study is to solve the limitations of slow convergence and untimely gradient update. Our specific contributions are summarized as follows:

- We propose QuAsyncFL, a new Asynchronous Federated Learning approach with Quantization for cloud-edge-terminal collaboration enabled AIoT. We adopt a new quantizer for natural compression to quantize the local gradients, where the quantizer is unbiased and nonuniform. The communication rounds can be reduced by quantizing the local gradients.
- We provide a detailed theoretical analysis on convergence speed. By utilizing the properties of convex functions and the unbiased and variance bounded of the quantizer, we prove that QuAsyncFL is convergent. As the number of iterations of the local computation increases, the upper bound of convergence becomes tighter.
- We conduct extensive performance evaluations of QuAsyncFL, demonstrating significant improvements over the original asynchronous federated learning and verifying the impact of quantification levels on communication rounds.

The remainder of this paper is organized as follows. In Section II, we present the system model. Next, we describe the detail design of the proposed QuAsyncFL method with theoretical analysis in Section III. Then, we evaluate QuAsyncFL's performance in Section IV. After reviewing the related work in Section V, we conclude our paper in Section VI, along with a discussion on future work.

II. System Model

This section discusses the system model in terms of network architecture, asynchronous federated learning, and gradient quantization.

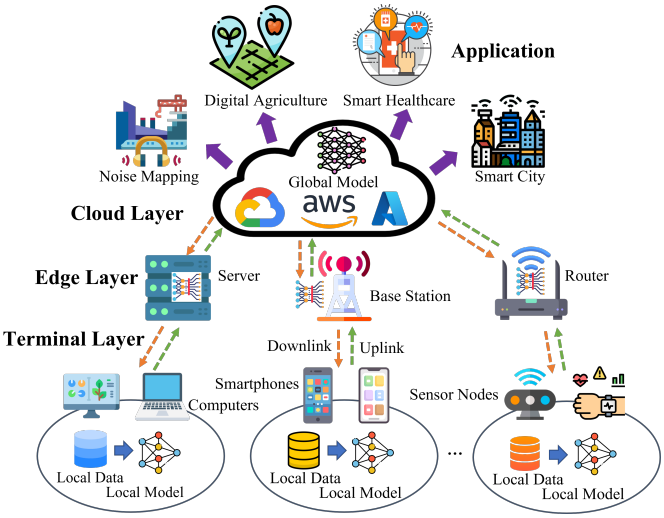


Fig. 1. Network architecture for AIoT with cloud-edge-terminal collaboration.

A. Network Architecture

This work considers a typical network architecture for AIoT, enabled by cloud-edge-terminal collaboration. The architecture is illustrated in Fig. 1, which consists of three main layers: the terminal layer, the edge layer, and the cloud layer.

The terminal layer comprises various smart devices, such as sensor nodes, smartphones, and computers, that have the dual responsibility of interacting with the physical world through sensing and actuation and training machine learning models. The edge layer is responsible for collecting and preprocessing data from the terminal devices and transmitting it to the cloud for further analysis. This layer typically includes edge servers, gateways, and base stations with more robust computing and storage capabilities for machine learning model training.

The cloud layer is responsible for storing and processing vast amounts of data generated by both the edge and terminal devices. It typically comprises a cloud computing platform, such as Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform (GCP), that offers a range of services, including data storage, compute resources, and machine learning tools. Furthermore, the cloud layer orchestrates the training process and aggregates the model updates from the edge and terminal devices to build global model.

Finally, this architecture enables a variety of smart applications, such as intelligent noise mapping [15] and sustainable digital agriculture [16], [17], among others, by leveraging the combined capabilities of the terminal, edge, and cloud layers.

B. Asynchronous Federated Learning

In this study, we explore a federated learning scenario involving N clients and a central server. In this context, clients refer to smart devices at the terminal layer, while the central server represents a cloud computing platform in the above network architecture. It should be noted that edge devices can be classified as either clients or central servers. Thus, the optimization goal of federated learning is to minimize the loss function, which can be formulated as follows:

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3290818

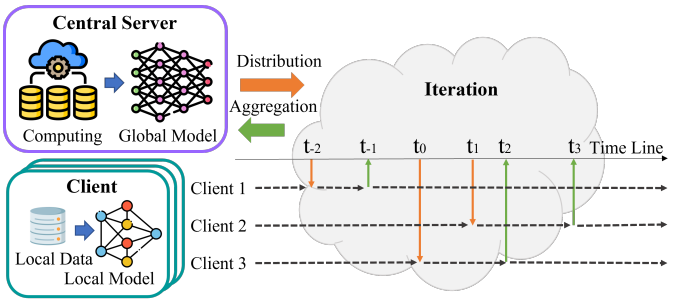


Fig. 2. Work procedure of asynchronous federated learning.

$$\min_{\mathbf{w}\in R^d} F(\mathbf{w}) = \frac{\sum_{i=1}^{N} D_i F_i(\mathbf{w})}{D} \tag{1}$$

where $F(\cdot)$ is the global loss function, $F_i(\cdot)$ is the local loss function, and $D = \sum_{i=1}^{N} D_i$ is the total datasets of all the clients.

Synchronous federated learning requires the central server to wait for all client nodes to complete local stochastic gradient descent (SGD) before performing the subsequent global parameter aggregation step [18]. As a result, communication efficiency is significantly reduced, and local devices may experience interruptions or long computing times, causing significant delays. To address these issues, we choose asynchronous federated learning, which enhances the flexibility of federated learning in cloud-edge-terminal collaboration enabled AIoT.

Specifically, we consider the typical work procedure of asynchronous federated learning [19], [20], as shown in Fig. 2. All participating clients independently perform their training process using their local data. Once a client has completed its local training for a predetermined number of epochs or iterations, it sends its updated model parameters to the central server, without waiting for other clients to complete their local training. At the central server side, it also does not need to wait for all clients to send their model updates before aggregating them, and can start performing the updates as soon as any of them arrives. After that, a new global model is calculated, and sent back to participating clients to update their local models.

C. Gradient Quantization

Gradient quantization is a technique used to compress multi-bit single-precision floating-point numbers into finite bits. Studies have shown that quantizing gradients can effectively reduce transmission bandwidth pressure and accelerate model training in SGD training of deep neural networks. The transmission framework model for this technique is illustrated in Fig. 3. After each iteration of gradient descent, the algorithm encodes and quantizes the gradient parameters, which are then transmitted to the next machine. In our work, the local clients transmit the quantized model parameters to the central server. On the server, the quantized parameters are decoded and used for global model aggregation.

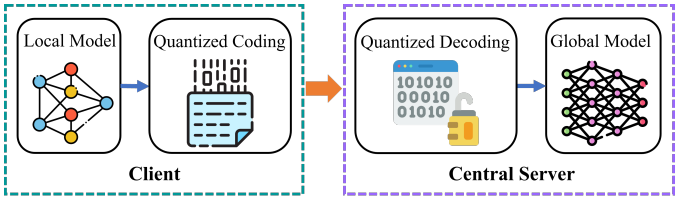


Fig. 3. Transmission framework of quantization stochastic gradient descent.

A low-precision quantizer for natural compression [21] is chosen for gradient quantization of local parameters. The quantizer is defined as:

$$Q(w_i) = \|\mathbf{w}\|_2 sgn(w_i)\mathcal{C}_i(w_i), \tag{2}$$

where

$$sgn(w_i) = \begin{cases} 1, & w_i > 0, \\ -1, & w_i \leq 0. \end{cases} \tag{3}$$

and

$$\mathcal{C}_i(w_i) = \begin{cases} 2^{\lfloor \log_2 |w_i| \rfloor}, & p(w_i), \\ 2^{\lceil \log_2 |w_i| \rceil}, & 1 - p(w_i). \end{cases} \tag{4}$$

where $p(w_i) = \frac{2^{\lceil \log_2 |w_i| \rceil} - |w_i|}{2^{\lfloor \log_2 |w_i| \rfloor}}$.

In multi-node communication transmission frameworks like federated learning, gradient quantization is a crucial to maintain the accuracy of the model while accelerating the model training speed. By compressing the gradients, it reduces the communication overhead between clients, that is essential for efficient and scalable distributed optimization.

III. QuAsyncFL Design

Building upon the system model discussed in the previous section, this section presents our proposed approach, QuAsyncFL, which is illustrated below.

A. QuAsyncFL

Similar to existing federated learning methods, the optimization goal of our proposed approach is to minimize the global loss function within a limited number of iterations and parameters. Specifically, we use stochastic gradient descent (SGD) to update the local model iteratively. To prevent overfitting during local model training, we employ gradient regularization. The parameters for the local single SGD iterative update on the i-th device are updated as follows:

$$\mathbf{w}_i(\tau) = \mathbf{w}_i(\tau - 1) - \eta \nabla G_i(\mathbf{w}_i(\tau - 1)), \tag{5}$$

where $G_i(\mathbf{w}_i(\tau)) = F_i(\mathbf{w}_i(\tau)) + \frac{\mu}{2}\|\mathbf{w}_i(\tau) - \mathbf{w}_i(0)\|_2^2$ and $G(\mathbf{w}) = \sum_{i=1}^{n} \frac{D_i G_i(\mathbf{w})}{D}$, $n \in \{1, \ldots, N\}$.

The global update method for asynchronous federated learning with quantization proposed here is described as follows:

$$\mathbf{w}(t) = (1 - \alpha)\mathbf{w}(t - 1) + \alpha Q(\mathbf{w}_{new}). \tag{6}$$

We complete the global aggregation of each round by assigning weights to the parameters generated by the previous aggregation round and the new parameters uploaded by local users, with the specific weights determined by different values

of α. Notably, $Q(\mathbf{w}_{new})$ refers to the parameters that have been quantized by natural compression in the local model.

Algorithm 1 Proposed Asynchronous Federated Learning with Quantization (QuAsyncFL)

Input: $\alpha \in (0,1)$
Initialize $\mathbf{w}_0$
Server Update:
for global aggregation $t = 1, 2, \ldots, k, \ldots, K$ do
Receive $(Q(\mathbf{w}_{new}), \tau)$ from the computing nodes
Update $\mathbf{w}(t) \leftarrow (1-\alpha)\mathbf{w}(t-1) + \alpha Q(\mathbf{w}_{new})$

Clients Update:
for each client $i = 1, 2, \ldots, N$ in parallel do
Receive $(\mathbf{w}(t), t)$ from the server
$k\tau \leftarrow t, \mathbf{w}_i(k\tau) \leftarrow \mathbf{w}(t)$
for local iteration $h = 1, 2, \ldots, H^i_{k\tau}$ do
Randomly sample $z^i_{k\tau,h}$ from D_i
Update $\mathbf{w}_i(k\tau + h) \leftarrow \mathbf{w}_i(k\tau + h - 1) - \eta \nabla G_i(\mathbf{w}_i(k\tau + h - 1); z^i_{k\tau,h})$
$Q(\mathbf{w}_{new}) \leftarrow \sum_{j=1}^{d} \|\mathbf{w}_i(k\tau+h)\|_2 sgn(w_{i,j}(k\tau+h)) \mathcal{C}_j(w_{i,j}(k\tau+h))$
Transmit $(Q(\mathbf{w}_{new}), \tau)$ to the server

The proposed QuAsyncFL design is a method that combines asynchronous federated learning and gradient quantization, with the addition of a gradient quantization step for local model parameters that does not compromise the flexibility of the federated learning model framework.

B. Quantization and Compression

In this paper, we explore a method called natural compression for quantizing parameters, which was discussed in Section II. The quantizer $Q(\cdot)$ provides an unbiased estimate, and the variance and L_2-norm of the parameter exhibit a positive correlation, as demonstrated in the definition of [21]:

Lemma 1: The quantization function satisfies the following properties:

$$(1)E[Q(\mathbf{w})] = \mathbf{w}$$
$$(2)E[\|Q(\mathbf{w}) - \mathbf{w}\|_2^2] \leq m\|\mathbf{w}\|_2^2,$$
$$(3)E[\|Q(\mathbf{w})\|_2^2 \leq (1+m)\|\mathbf{w}\|_2^2,$$

and m=$\frac{1}{8} + \min(\frac{\sqrt{d}}{2^{s-1}}, \frac{d}{2^{2(s-1)}})$, according to the Theorem 7 in [21].

For the sake of computational convenience, we propose two assumptions:

Assumption 1: $F_i(\mathbf{w})$ is convex, and F has Lipschitz continuous gradients with constant $\rho \geq 0$: $F_i(\mathbf{w}) - F_i(\mathbf{w}') \leq \langle \nabla F_i(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle + \frac{\rho}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2$.

Assumption 2: F and G is a differentiable function,$\forall \mathbf{w} \in R^d$,$\|\nabla F_i(\mathbf{w})\|_2 \leq V_1$,$\|\nabla G_i(\mathbf{w})\|_2 \leq V_2$, where V_1 and V_2 is a constant.

By applying natural compression to the local model, we observe a gap between the parameters from two consecutive rounds of global aggregation.

Lemma 2: We assume that when $t = k\tau$, the local device sends its parameter to the base station. If the i-th device receives the global parameter and updates h rounds, and $h \in [0, H]$, $H \in [H_{min}, H_{max}]$, we have

$$\begin{aligned} &E[F(\mathbf{w}_i(k\tau)) - F(\mathbf{w}(t-1))] \\ &\leq \frac{1}{2}\rho H_{max} A^2 + \sqrt{V_1}A + \frac{3}{2}H_{max}B^2 + \sqrt{V_1}B + \\ &\frac{3}{2}\rho H_{max} C w_{im}^2 + (1+\alpha)\sqrt{V_1}w_{im}, \end{aligned} \tag{7}$$

where A=$\eta H_{max}\sqrt{V_2}$, B=$(1-\alpha)w_m$, C=$(1+\alpha^2(1+m))$, w_{im}=$\max\{\|\mathbf{w}_i(k\tau+h)\|_2\}$ and w_m=$\max\{\|\mathbf{w}(t)\|_2\}$, H_{max} is the max local iteration in the clients, k is the index of aggregation.

Proof: See Appendix A. ■

C. Convergence Analysis

To facilitate convenient convergence analysis and derivation, we derive several related lemmas and theorems. From the Assumption 1, F is a convex function, then we have

Lemma 3: $F_i(\mathbf{w})$,$G_i(\mathbf{w})$ and $G(\mathbf{w})$ is also convex,ρ-Lipschitz and β-smooth.

Lemma 4: Suppose F satisfies Assumption 1 and 2, then the bound of initial loss on $\mathbf{w}_i(k\tau + h)$ is given by

$$\begin{aligned} &F(\mathbf{w}_i(k\tau+h)) - F(\mathbf{w}_i(k\tau+h-1)) \\ &\leq -\eta\frac{nD_{max}}{D}\sum_{i=1}^{n}\|\nabla F_i(\mathbf{w}_i(k\tau+h-1))\|_2^2 + \\ &\frac{1}{2}\mu\eta^2 H_{max}^2 V_2 + \frac{1}{2}\eta^2\rho V_2. \end{aligned} \tag{8}$$

Proof: See Appendix B. ■

Based on the lemma and assumption, we can derive several theorems for convergence analysis of QuAsyncFL. Thus, we need to derive a definite upper bound for the loss function of two consecutive rounds, which is given by

Theorem 1: The upper bound of the gap is given as follows

$$\begin{aligned} &E[F(\mathbf{w}(t)) - F(\mathbf{w}(t-1))] \\ &\leq -\alpha\eta\frac{nD_{max}}{D}\sum_{h}^{H-1}\sum_{i=1}^{n}\|\nabla F_i(\mathbf{w}_i(k\tau+h))\|_2^2 + \frac{1}{2}\eta^2\alpha\rho V_2 + \\ &\frac{1}{2}\alpha\rho H_{max}A^2 + \alpha\sqrt{V_1}A + 3\alpha B^2(\frac{1}{2}H_{max} + \mu) + \alpha\sqrt{V_1}B + \\ &\alpha w_{im}((1+\alpha)\sqrt{V_1} + \mu m + \frac{3}{2}H_{max}C) + 3\alpha\mu C w_{im}^2. \end{aligned} \tag{9}$$

Proof: See Appendix C. ■

By utilizing the upper bound discussed in Theorem 1, we establish a critical point of convergence.

Theorem 2: QuAsyncFL converges to a critical point:

$$\frac{1}{\sum_{k=0}^{K} H_k} \sum_{k=1}^{K} E[\|\nabla F(\mathbf{w}(t))\|_2^2] \leq O(\frac{\delta}{\epsilon H_{min}^2}) + O(\frac{1}{\epsilon H_{min}^3}) + O(\frac{\delta^3}{\epsilon H_{min}^4}) + O(\frac{1}{\epsilon H_{min}^5}) + O(\frac{\delta}{\epsilon H_{min}^6}), \quad (10)$$

where $\delta = \frac{H_{max}}{H_{min}}$, $\epsilon = \frac{nD_{max}}{D}$, n is the number of users participating in the current round of aggregation and $n \in \{1, \ldots, N\}$, K is the total rounds of aggregation, H_k is the rounds of local training after the $(k-1)$-th aggregation.

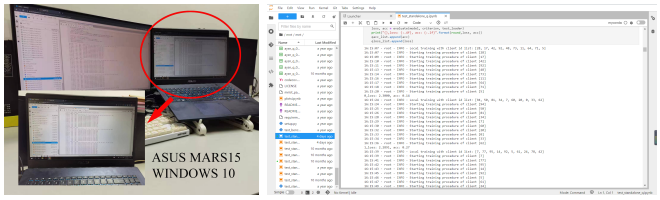
Proof: See Appendix D. ∎

Through the critical point in Theorem 2, we can find that the algorithm has a clear convergence boundary, that is, as the number of iterations increases, the convergence boundary of the algorithm tends to 0, and the loss function will be infinitely close to the minimum value, which means the algorithm can reach a convergent state.

IV. Performance Evaluation

This section presents the performance evaluation of QuAsyncFL, which was conducted using the PyTorch-based framework provided by FedLab [22]. The experiment was conducted using an ASUS MARS15 laptop running Windows 10, while the program was executed on Matpool with an NVIDIA Tesla K80, as depicted in Fig. 4.

We employed the CNN model on the clients and utilized the MNIST dataset [23] as the training set. The relevant parameter settings of the CNN model are displayed in Table I. We aslo set the number of users to 100 and the batch size to 100. For local training, we trained each client's model for 5 epochs with a learning rate of 0.02 using stochastic gradient descent. The loss function we used was the cross-entropy loss function.

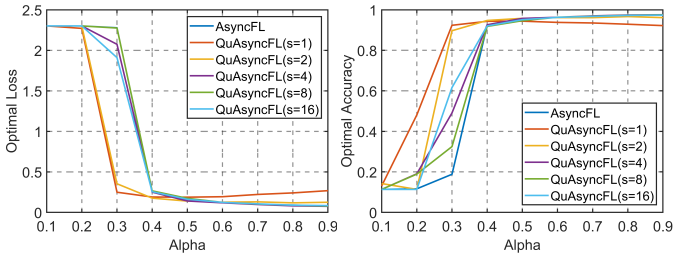


(a) Laptop (b) Program running interface

Fig. 4. Snapshot of experimental environment for performance evaluation.

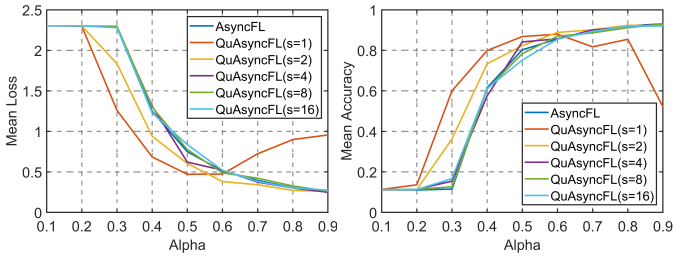
TABLE I
THE RELEVANT PARAMETERS SETTING

Type	Dimension	Size	Step Size	Activation
Input	1	-	-	-
Conv	32	5*5	-	-
MaxPool	32	2*2	1	-
Conv	64	5*5	-	-
MaxPool	64	2*2	1	-
Linear	512	-	-	ReLU
Linear	10	-	-	-



(a) Optimal loss (b) Optimal accuracy

Fig. 5. The results of optimal values by testing α from 0.1 to 0.9 at the non-quantization and the quantization levels s of 1, 2, 4, 8, and 16, respectively.



(a) Mean loss (b) Mean accuracy

Fig. 6. The results of mean values by testing α from 0.1 to 0.9 at the non-quantization and the quantization levels s of 1, 2, 4, 8, and 16, respectively.

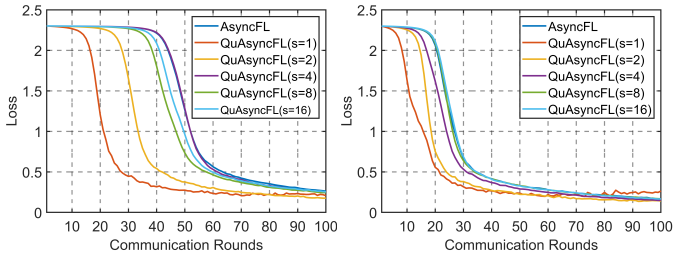
A. Parameters Adjustment

First of all, we searched for the optimal value of α by gradually testing values from 0.1 to 0.9. Results from Fig. 5 show that model performance was unsatisfactory before α was set to 0.4. Even after the same number of communications, the loss function's value continued to converge at a higher level, and the model's accuracy was consistently poor. While model performance improved after 0.4, QuAsyncFL with s=1 showed training performance rebounding after 0.6, indicating overfitting. For example, when α=0.7, the mean accuracy decreases by 10% compared to 0.6. Meanwhile, results from Fig. 6 show that the mean loss trended downward and the mean accuracy upward after α was set to 0.2, but rebounded for α values beyond 0.6. Therefore, model training performance was best when α was set to 0.4 or 0.5. Based on these results, we chose to use α values of 0.4 and 0.5 for preliminary training.

B. Communication Efficiency

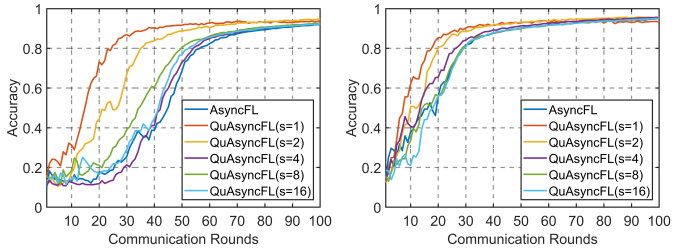
After that, we conducted an evaluation of the communication efficiency of asynchronous federated learning, comparing it with and without quantization. The results demonstrate that higher quantization levels lead to fewer communication rounds required for convergence. As shown in Fig. 7, the convergence rate of QuAsyncFL with s=1 is 75% faster than the non-quantization method. However, as the quantization level decreases, the number of communication rounds increases, which requires more parameter aggregation and transmission during training. The convergence trend with a higher quantization level is similar to that of the non-quantization framework. Therefore, there exists an upper bound for s in the quantization process.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3290818



(a) α=0.4 (b) α=0.5

Fig. 7. The results of loss versus communication rounds.



(a) α=0.4 (b) α=0.5

Fig. 8. The results of accuracy versus communication rounds.

C. Accuracy

Meanwhile, we also evaluated the accuracy of asynchronous federated learning with and without quantization. The results showed some interesting observations, as demonstrated in Fig. 8. We found that QuAsyncFL with s=1 model converges very fast in a defined number of communications, with the accuracy reaching 12.5% higher than the non-quantization model around the 50th aggregation. However, as the quantization level increases, the overfitting problem becomes more prominent. Due to the large gradient, the model reaches the optimal value quickly. However, as the training time increases, the model becomes over-trained, and the original regularization method is not sufficient to reduce the risk of overfitting. As a result, the performance of QuAsyncFL with s=1 model is not as good as the non-quantization model or QuAsyncFL with a lower quantization level.

D. Performance under Different Weights

Finally, we discuss the impact of different aggregation weights on the performance of QuAsyncFL. By comparing the results in Fig. 7 and Fig. 8, we found that different aggregation weights (i.e., different α values) can significantly affect the convergence rate. When s=1 and s=2, there is no significant difference in the convergence rate. However, as the quantization level decreases, the impact of the aggregation weight becomes more evident. For instance, when s=4, QuAsyncFL with α=0.5 is 1.8x faster than QuAsyncFL with α=0.4. Similarly, the non-quantization model with α=0.5 converges 1.71x faster than with α=0.4. It is worth noting that using a higher aggregation weight α may increase the risk of overfitting during the training process. Therefore, the choice of aggregation weight in QuAsyncFL should be based on the specific quantization level used.

V. Related Work

This section analyses the related work of our study with respect to the following three domains and Table II.

Cloud-Edge-Terminal Collaboration. It has emerged as a promising approach for achieving efficient and scalable IoT systems. In recent years, a large and growing body of literature has investigated this area [24], [25]. For example, one advancement is the use of a blockchain-assisted collective Q-learning approach for networking integrated cloud-edge-end in IoT [26]. Another promising approach is the use of federated learning frameworks for distributed deep neural networks over cloud, edge, and end devices [27]. Several works have also proposed cloud-edge-terminal collaboration for specific applications such as temperature measurement in COVID-19 prevention [28], wind turbine damage detection [29], virtual reality [30], and more [31], [32]. These applications demonstrate the potential of cloud-edge-end collaboration for improving efficiency and reducing costs in various domains.

While there is a growing interest in this topic, there is still a need for research on the trade-offs between computational efficiency, network bandwidth, and data privacy in cloud-edge-end collaboration. Our work focuses on asynchronous federated learning, which enables efficient collaborative learning without the need for centralized data storage or processing, thereby enabling privacy preservation and reducing network latency. We select an aggregation model with better performance by adjusting the aggregated weight in aggregation to improve flexibility of federated learning. Simultaneously, the natural compression method is combined to quantize the local parameters to improve the communication efficiency of asynchronous federated learning. This enables asynchronous federated learning to effectively protect data security in AIoT applications, while also improving the flexibility and efficiency of communication.

Asynchronous Federated Learning. This direction has become increasingly popular in recent years due to its ability to handle large-scale and heterogeneous datasets [33]. A significant amount of literature has been published, with a focus on its applications, mechanisms, and performance enhancements. For example, various studies have investigated the potential of asynchronous federated learning in edge computing [34], IoT [35], vehicular networks [36], fault diagnosis [37], and critical energy infrastructure [38]. Adaptive methods have been proposed for asynchronous federated learning in resource-constrained edge computing [39], while scheduling and aggregation methods have been developed to improve its performance over wireless networks [40].

Notably, gradient quantization has been recently explored in a few studies to enhance the performance of federated learning. For instance, adaptive gradient quantization has been proposed to achieve communication-efficient federated learning in heterogeneous edge devices [41]. A method was proposed to improve the model prediction accuracy with the system latency guarantee and gradient quantization in federated learning [42]. An asynchronous federated learning method has been presented to leverage majority voting to combine quantized model updates from different edge devices [43].

TABLE II
THE REFERENCE OF RELATED WORK

Reference	Cloud-Edge-Terminal Collaboration	Gradient Quantization	Synchronous Federated Learning	Asynchronous Federated Learning
[24]–[26], [28]–[32]	✓	-	-	-
[27]	✓	-	-	✓
[33]–[40]	-	✓	-	-
[41], [42], [44], [45], [47]	-	-	✓	✓
[43]	-	✓	✓	-
[46], [48]–[50]	-	-	✓	-
Our work	✓	✓	-	✓

Moreover, lazily quantized gradient [44] and heterogeneous quantization [45] have been proposed for dynamic aggregation in federated learning. However, our study differs from these works in several ways. Firstly, we introduce an innovative approach by adjusting the aggregated weight in aggregation to select an aggregation model with better performance, thereby improving the flexibility of federated learning. Secondly, we combine the natural compression method with our approach to quantize the local parameters, which significantly improves the communication efficiency of asynchronous federated learning. As a result, our proposed approach effectively protects data security in AIoT applications while improving the flexibility and efficiency of communication.

Gradient Quantization. Finally, we provide a brief overview of the related work on gradient quantization. This technique has been successfully applied in various domains such as distributed deep learning [46], mobile and edge computing [47], natural language processing [48], and computer vision [49]. Gradient quantization is particularly useful in reducing the memory footprint and communication overhead during distributed training [50], making it possible to train larger models in distributed environments.

In our study, we take advantage of a low-precision quantizer to perform gradient quantization of local parameters. By doing so, we improve the efficiency of cloud-edge-terminal collaboration for AIoT applications, where resources are often limited. This allows us to take full advantage of the distributed computing capabilities and improve the efficiency of services.

VI. CONCLUSION AND FUTURE WORK

Conclusion. In this paper, we proposed QuAsyncFL, a quantized federated learning framework for improving the communication efficiency of the system without reducing the flexibility of asynchronous federated learning. Our focus is on applying this approach to cloud-edge-terminal collaboration enabled AIoT. QuAsyncFL adopts weight aggregation for the model aggregation method of asynchronous federated learning and combines the natural compression method for quantizing the local parameters, thereby improving the communication efficiency of distributed learning with limited resources. Our experiments present the convergence difference between QuAsyncFL with asynchronous federated learning and show that different quantization levels have a certain impact on the convergence of the model.

Limitations and Future Work. While QuAsyncFL improves communication efficiency, it still has some limitations. Quantization inevitably brings quantization errors, which can be more significant with greater quantization. In addition, individual clients may interrupt contact due to poor wireless signal quality, which can impact service quality. To address these issues, future work could investigate relevant factors for improving communication efficiency without excessive quantization error, explore other quantizers that can reduce the generation of quantization error, or consider collaborative learning to enable disconnected clients to still participate in training. Other solutions could include introducing reference factors in cloud-edge-terminal collaboration scenario to reduce the probability of client disconnection.

APPENDIX A
PROOF OF LEMMA 2

For the convenience of calculation, we first derive the upper bound of the L_2-norm of the gap, where the gap is the parameters of two consecutive rounds in the global aggregation.

Assume that when $t = k\tau$, the local device sends its parameter to the server. If the i-th device receives the global parameter and updates h rounds, and $h \in [0, H]$,$H \in [H_{min}, H_{max}]$, we derive an upper bound with using Assumption 1 is given by

$$
\begin{aligned}
&E[\|\mathbf{w}_i(k\tau) - \mathbf{w}(t-1)\|_2^2] \\
&= E[\|\mathbf{w}_i(k\tau) - \mathbf{w}_i(k\tau+1) + \mathbf{w}_i(k\tau+1) - \mathbf{w}_i(k\tau+2) + \cdots \\
&+ \mathbf{w}_i(k\tau+H-1) - \mathbf{w}_i(k\tau+H) + \mathbf{w}_i(k\tau+H) - \mathbf{w}(t-1)\|_2^2] \\
&\overset{(a)}{\leq} E[H_{max}\|\mathbf{w}_i(k\tau) - \mathbf{w}_i(k\tau+1)\|_2^2 + H_{max}\|\mathbf{w}_i(k\tau+1) - \\
&\mathbf{w}_i(k\tau+2)\|_2^2 + \cdots + H_{max}\|\mathbf{w}_i(k\tau+H-1) - \\
&\mathbf{w}_i(k\tau+H)\|_2^2 + H_{max}\|\mathbf{w}_i(k\tau+H) - \mathbf{w}(t-1)\|_2^2] \\
&\leq H_{max}^3\eta^2 V_2 + E[H_{max}\|\mathbf{w}_i(k\tau+H) - \mathbf{w}(t-1)\|_2^2] \\
&= H_{max}^3\eta^2 V_2 + H_{max}E[\|\mathbf{w}_i(k\tau+H) - ((1-\alpha)\mathbf{w}(t-2) + \\
&\alpha Q(\mathbf{w}_i(k\tau+H)))\|_2^2] \\
&\overset{(a)}{\leq} H_{max}^3\eta^2 V_2 + H_{max}E[3\|\mathbf{w}_i(k\tau+H)\|_2^2 + \\
&3\|\alpha Q(\mathbf{w}_i(k\tau+H))\|_2^2 + 3\|(1-\alpha)\mathbf{w}(t-2)\|_2^2] \\
&\overset{(b)}{\leq} H_{max}^3\eta^2 V_2 + 3H_{max}E[\|\mathbf{w}_i(k\tau+H)\|_2^2 + \alpha^2(1+ \\
&m)\|\mathbf{w}_i(k\tau+H)\|_2^2 + (1-\alpha)^2\|\mathbf{w}(t-2)\|_2^2] \\
&\leq H_{max}^3\eta^2 V_2 + 3H_{max}(1+\alpha^2(1+m))w_{im}^2 + 3H_{max}(1-\alpha)^2 w_m^2,
\end{aligned}
\tag{11}
$$

where the inequality (a) is based on $\|\Sigma_{i=1}^{N} x_i\|_2^2 \leq N\Sigma_{i=1}^{N}\|x_i\|_2^2$, the inequality (b) is based on Lemma 1. With the same. With the similar derivation, we also have

$$\begin{aligned}
&E[\|\mathbf{w}_i(k\tau) - \mathbf{w}(t-1)\|_2] \\
&= E[\|\mathbf{w}_i(k\tau) - \mathbf{w}_i(k\tau+1) + \mathbf{w}_i(k\tau+1) - \mathbf{w}_i(k\tau+2) + \cdots \\
&+ \mathbf{w}_i(k\tau+H-1) - \mathbf{w}_i(k\tau+H) + \mathbf{w}_i(k\tau+H) - \mathbf{w}(t-1)\|_2] \\
&\overset{(c)}{\leq} E[\|\mathbf{w}_i(k\tau)\mathbf{w}_i(k\tau+1)\|_2 + \cdots + \|\mathbf{w}_i(k\tau+H) - w(t-1)\|_2] \\
&\leq H_{max}\eta\sqrt{V_2} + E[\|\mathbf{w}_i(k\tau+H) - (1-\alpha)\mathbf{w}(t-2) - \\
&\alpha Q(\mathbf{w}_i(k\tau+H))\|_2] \\
&\overset{(c)}{\leq} H_{max}\eta\sqrt{V_2} + E[\|\mathbf{w}_i(k\tau+H)\|_2 + \|(1-\alpha)\mathbf{w}(t-2)\|_2 + \\
&\|\alpha Q(\mathbf{w}_i(k\tau+H))\|_2] \\
&= H_{max}\eta\sqrt{V_2} + [(1+\alpha)\|\mathbf{w}_i(k\tau+H)\|_2 + (1-\alpha)\|\mathbf{w}(t-2)\|_2] \\
&\leq H_{max}\eta\sqrt{V_2} + (1+\alpha)\ w_{im} + (1-\alpha)w_m,
\end{aligned} \tag{12}$$

where the inequality (c) is based on the triangle inequality, w_{im}=$\max\{\|\mathbf{w}_i(k\tau+h)\|_2\}$ and w_m=$\max\{\|\mathbf{w}(t)\|_2\}$. Using (11) and (12), we can derive a gap between the parameters from two consecutive global aggregations as follows

$$\begin{aligned}
&E[F(\mathbf{w}_i(k\tau)) - F(\mathbf{w}(t-1))] \\
&\leq \langle \nabla F(\mathbf{w}(t-1)), \mathbf{w}_i(k\tau) - w(t-1)\rangle + \frac{\rho}{2}\|\mathbf{w}_i(k\tau) - \mathbf{w}(t-1)\|_2^2 \\
&\leq \|\nabla F(\mathbf{w}(t-1))\|_2\|\mathbf{w}_i(k\tau) - \mathbf{w}(t-1)\|_2 + \frac{\rho}{2}\|\mathbf{w}_i(k\tau) - \mathbf{w}(t-1)\|_2^2 \\
&\leq \sqrt{V_1}\ (\eta H_{max}\sqrt{V_2} + (1+\alpha)w_{im} + (1-\alpha)w_m) + \\
&\frac{\rho}{2}H_{max}(\eta^2 H_{max}^2 V_2 + 3(1+\alpha^2(1+m))w_{im}^2 + 3(1-\alpha)^2 w_m^2) \\
&= \frac{1}{2}\rho H_{max}A^2 + \sqrt{V_1}A + \frac{3}{2}H_{max}B^2 + \sqrt{V_1}B + \frac{3}{2}\rho H_{max}C w_{im}^2 \\
&+ (1+\alpha)\sqrt{V_1}w_{im}.
\end{aligned} \tag{13}$$

where A=$\eta H_{max}\sqrt{V_2}$, B=$(1-\alpha)w_m$, C=$(1+\alpha^2(1+m))$. This completes the proof.

APPENDIX B
PROOF OF LEMMA4

Similarly, with using hypothesis in the proof of Lemma 2, Assumption 1 and 2, we derive an upper bound of initial loss on $w_i(k\tau+h)$ is given by

$$\begin{aligned}
&E[F(\mathbf{w}_i(k\tau+h)) - F(\mathbf{w}^*)] \\
&\overset{(d)}{\leq} E[G(\mathbf{w}_i(k\tau+h)) - F(\mathbf{w}^*)] \\
&= E[G(\mathbf{w}_i(k\tau+h)) - G(\mathbf{w}_i(k\tau+h-1)) \\
&+ G(\mathbf{w}_i(k\tau+h-1)) - F(\mathbf{w}^*)],
\end{aligned} \tag{14}$$

where the inequality (d) follows $G(\mathbf{w}_i(k\tau+h))$ = $F(\mathbf{w}_i(k\tau+h)) + \frac{\mu}{2}\|\mathbf{w}_i(k\tau+h) - \mathbf{w}_i(k\tau)\|_2^2$. With Assumption 1, we have

$$\begin{aligned}
&G(\mathbf{w}_i\ (k\tau+h)) - G(\mathbf{w}_i\ (k\tau+h-1)) \\
&\leq \langle \nabla G(\mathbf{w}_i(k\tau+h-1)), \mathbf{w}_i(k\tau+h) - \mathbf{w}_i(k\tau+h-1)\rangle + \\
&\frac{\rho}{2}\|\mathbf{w}_i(k\tau+h) - \mathbf{w}_i(k\tau+h-1)\|_2^2 \\
&= \langle \nabla G(\mathbf{w}_i(k\tau+h-1)), -\eta\nabla G_i(\mathbf{w}_i(k\tau+h-1))\rangle + \\
&\frac{\rho}{2}\eta^2\|\nabla G_i(\mathbf{w}_i(k\tau+h-1))\|_2^2.
\end{aligned} \tag{15}$$

Rearrange the formula as follows:

$$\begin{aligned}
&\langle \nabla G(\mathbf{w}_i(k\tau+h-1)), \nabla G_i(\mathbf{w}_i(k\tau+h-1))\rangle \\
&= \left\langle \nabla \frac{\sum_{i=1}^{n} D_i G_i(\mathbf{w}_i(k\tau+h-1))}{D}, \nabla G_i(\mathbf{w}_i(k\tau+h-1)) \right\rangle \\
&\leq \frac{nD_{max}}{D}\sum_{i=1}^{n}\langle \nabla G_i(\mathbf{w}_i(k\tau+h-1)), \nabla G_i(\mathbf{w}_i(k\tau+h-1))\rangle \\
&= \frac{nD_{max}}{D}\sum_{i=1}^{n}\|\nabla G_i(\mathbf{w}_i(k\tau+h-1))\|_2^2,
\end{aligned} \tag{16}$$

where k is the index of aggregation and D_{max} is the largest D_i of all clients. According to the realation between $G(\mathbf{w})$ and $F(\mathbf{w})$, we have $\frac{nD_{max}}{D}\sum_{i=1}^{n}\|\nabla G_i(\mathbf{w}_i(k\tau+h-1))\|^2 \geq \frac{nD_{max}}{D}\sum_{i=1}^{n}\|\nabla F_i(\mathbf{w}_i(k\tau+h-1))\|^2$. Based on the inequality (d) and (16), we get an inequality given by

$$\begin{aligned}
&-\eta\frac{nD_{max}}{D}\sum_{i=1}^{n}\|\nabla G_i(\mathbf{w}_i(k\tau+h-1))\|_2^2 \\
&\leq -\eta\frac{nD_{max}}{D}\sum_{i=1}^{n}\|\nabla F_i(\mathbf{w}_i(k\tau+h-1))\|_2^2.
\end{aligned} \tag{17}$$

By rearranging the (14), (15) and (17), we have:

$$\begin{aligned}
&E[F(\mathbf{w}_i\ (k\tau+h)) - F(\mathbf{w}^*)] \\
&\leq E[G(\mathbf{w}_i(k\tau+h)) - G(\mathbf{w}_i(k\tau+h-1)) + \\
&G(\mathbf{w}_i(k\tau+h-1)) - F(\mathbf{w}^*)] \\
&\leq E[\langle \nabla G(\mathbf{w}_i(k\tau+h-1)), -\eta\nabla G_i(\mathbf{w}_i(k\tau+h-1))\rangle + \\
&\frac{\eta^2\rho}{2}\|\nabla G_i(\mathbf{w}_i(k\tau+h-1))\|_2^2 + G(\mathbf{w}_i(k\tau+h-1)) - F(\mathbf{w}^*)] \\
&= \langle \nabla G(\mathbf{w}_i(k\tau+h-1)), -\eta\nabla G_i(\mathbf{w}_i(k\tau+h-1))\rangle + \\
&\frac{\eta^2\rho}{2}\|\nabla G_i(\mathbf{w}_i(k\tau+h-1))\|_2^2 + F(\mathbf{w}_i(k\tau+h-1)) + \\
&\frac{\mu}{2}\|\mathbf{w}_i(k\tau+h-1) - \mathbf{w}_i(k\tau)\|_2^2 - F(\mathbf{w}^*) \\
&\leq F(\mathbf{w}_i(k\tau+h-1)) - F(\mathbf{w}^*) + \frac{1}{2}\mu\eta^2 H_{max}^2 V_2 + \frac{\eta^2\rho}{2}V_2 - \\
&\eta\frac{nD_{max}}{D}\sum_{i=1}^{n}\|\nabla F_i(\mathbf{w}_i(k\tau+h-1))\|_2^2.
\end{aligned} \tag{18}$$

By rearranging (18), we get an upper bound given by

$$\begin{aligned}
&F(\mathbf{w}_i(k\tau+h)) - F(\mathbf{w}_i(k\tau+h-1)) \\
&\leq -\eta\frac{nD_{max}}{D}\sum_{i=1}^{n}\|\nabla F_i(\mathbf{w}_i(k\tau+h-1))\|_2^2 + \\
&\frac{1}{2}\mu\eta^2 H_{max}^2 V_2 + \frac{1}{2}\eta^2\rho V_2.
\end{aligned} \tag{19}$$

This completes the proof.

APPENDIX C
PROOF OF THEOREM1

Using the above lemmas, we can gain a gap of the parameters of two consecutive rounds after aggregation. The upper

bound of the gap is given as follows

$$
\begin{aligned}
&E[F(\mathbf{w}(t)) - F(\mathbf{w}(t-1))] \\
&\leq E[G(\mathbf{w}(t)) - F(\mathbf{w}(t-1))] \\
&\leq E[(1-\alpha)G(\mathbf{w}(t-1)) + \alpha G(Q(\mathbf{w}_i(k\tau+H)) - F(\mathbf{w}(t-1))] \\
&\overset{(e)}{=} E[-\alpha F(\mathbf{w}(t-1)) + \alpha G(Q(\mathbf{w}_i(k\tau+H))] \\
&= E[-\alpha F(\mathbf{w}(t-1)) + \alpha F(Q(\mathbf{w}_i(k\tau+H)) + \\
&\frac{\alpha\mu}{2}\|Q(\mathbf{w}_i(k\tau+H)) - \mathbf{w}(t-1)\|_2^2] \\
&= \alpha E[-F(\mathbf{w}(t-1)) - F(\mathbf{w}_i(k\tau)) + F(\mathbf{w}_i(k\tau)) + \\
&F(Q(\mathbf{w}_i(k\tau+H))) + \frac{\mu}{2}\|Q(\mathbf{w}_i(k\tau+H)) - \mathbf{w}_i(k\tau+H) + \\
&\mathbf{w}_i(k\tau+H) - \mathbf{w}(t-1)\|_2^2] \\
&\leq \alpha E[F(\mathbf{w}_i(k\tau)) - F(\mathbf{w}(t-1)) + F(Q(\mathbf{w}_i(k\tau+H)) - \\
&F(\mathbf{w}_i(k\tau)) + \mu\|Q(\mathbf{w}_i(k\tau+H)) - \mathbf{w}_i(k\tau+H)\|_2^2 + \\
&\mu\|\mathbf{w}_i(k\tau+H) - \mathbf{w}(t-1)\|_2^2] \\
&\leq -\alpha\eta\frac{nD_{max}}{D}\sum_{h}^{H-1}\sum_{i=1}^{n}\|\nabla F_i(\mathbf{w}_i(k\tau+h))\|_2^2 + \alpha\eta H_{max}(\sqrt{V_1V_2} + \\
&\frac{\mu}{2}H_{max}^2V_2) + \alpha w_{im}^2(\frac{3}{2}\rho H_{max}(1+\alpha^2(1+m)) + \mu(m+(1-\alpha)^2)) \\
&+ \alpha w_m^2(\frac{3}{2}\rho H_{max}(1-\alpha^2) + \mu(1-\alpha)^2) + \alpha\sqrt{V_1}((1+\alpha)w_{im} + \\
&(1-\alpha)w_m) + \frac{1}{2}\eta^2\alpha\rho H_{max}V_2(H_{max}^2+1) \\
&= -\alpha\eta\frac{nD_{max}}{D}\sum_{h}^{H-1}\sum_{i=1}^{n}\|\nabla F_i(\mathbf{w}_i(k\tau+h))\|_2^2 + \frac{1}{2}\eta^2\alpha\rho V_2 + \\
&\frac{1}{2}\alpha\rho H_{max}A^2 + \alpha\sqrt{V_1}A + 3\alpha B^2(\frac{1}{2}H_{max} + \mu) + \alpha\sqrt{V_1}B + \\
&\alpha w_{im}((1+\alpha)\sqrt{V_1} + \mu m + \frac{3}{2}H_{max}C) + 3\alpha\mu C w_{im}^2,
\end{aligned}
\tag{20}
$$

where equality (e) is based on $G(\mathbf{w}(t-1)) = F(\mathbf{w}(t-1)) - \frac{\mu}{2}\|\mathbf{w}(t-1) - \mathbf{w}(t-1)\|_2^2$. This completes the proof.

Appendix D
Proof of Theorem2

From Theorem 2, we can get a critical point as follows

$$
\begin{aligned}
&\sum_{h}^{H-1}\sum_{i=1}^{n}\|\nabla F_i(\mathbf{w}_i(k\tau+h))\|_2^2 \\
&\leq \frac{1}{\alpha\eta\epsilon}[F(\mathbf{w}(t)) - F(\mathbf{w}(t-1))] + \frac{1}{\eta\epsilon}\sqrt{V_1}(\eta H_{max}\sqrt{V_2} + \\
&(1+\alpha)w_{im} + (1-\alpha)w_m) + \frac{1}{2\eta\epsilon}\rho H_{max}(\eta^2H_{max}^2V_2 \\
&+ 3(1+\alpha^2(1+m))w_{im}^2 + 3(1-\alpha)^2w_m^2) + \frac{1}{2\epsilon}\rho\eta H_{max}V_2 \\
&+ \frac{1}{2\epsilon}\alpha\mu H_{max}^3V_2 + \frac{1}{\eta\epsilon}\mu m w_{im}^2 + \frac{1}{\eta\epsilon}\mu((1+\alpha^2(1+m))w_{im}^2 \\
&+ (1-\alpha)^2w_m^2),
\end{aligned}
\tag{21}
$$

where $\epsilon = \frac{nD_{max}}{D}$.

By rearranging (21), an upper bound of the critical point is given by

$$
\begin{aligned}
&\frac{1}{\sum_{k=0}^{K}H_k}\sum_{k=1}^{K}E[\|\nabla F(\mathbf{w}(t))\|_2^2] \\
&\leq \frac{1}{\epsilon K H_{min}}(\frac{1}{\alpha\eta}[F(\mathbf{w}(0)) - F(\mathbf{w}(K))] + \frac{1}{\eta}(\eta H_{max}\sqrt{V_2} + \\
&(1+\alpha)w_{im} + (1-\alpha)w_m) + \frac{1}{2\eta}\rho H_{max}(\eta^2H_{max}^2V_2 \\
&+ 3(1+\alpha^2(1+m))w_{im}^2 + 3(1-\alpha)^2w_m^2) + \frac{1}{2}\rho\eta H_{max}V_2 + \\
&\frac{1}{2}\alpha\mu H_{max}^3V_2 + \frac{1}{\eta}\mu m w_{im}^2 + \frac{1}{\eta}\mu((1+\alpha^2(1+m))w_{im}^2 \\
&+ (1-\alpha)^2w_m^2)),
\end{aligned}
\tag{22}
$$

where K is the total rounds of aggregation, H_k is the rounds of local training after the $(k-1)$-th aggregation.

Taking $K = H_{min}^4$,$\eta = \frac{1}{\sqrt{K}} = \frac{1}{H_{min}^2}$,$\alpha = \frac{1}{H_{min}}$,$\delta = \frac{H_{max}}{H_{min}}$, we have

$$
\begin{aligned}
&\frac{1}{\sum_{k=0}^{K}H_k}\sum_{k=1}^{K}E[\|\nabla F(\mathbf{w}(t))\|_2^2] \\
&\leq \frac{1}{\epsilon H_{min}^2}[(F(\mathbf{w}(0)) - F(w(K))) + \frac{3}{2}\delta\rho w_{im}^2 + \frac{3}{2}\rho w_m^2] + \\
&\frac{1}{\epsilon H_{min}^3}[\sqrt{V_1}w_{im} + \sqrt{V_1}w_m - 3\rho w_m^2 + \frac{1}{2}V_2 + \mu m w_{im}^2 + \\
&\mu w_{im}^2 + \mu w_{im}^2] + \frac{1}{\epsilon H_{min}^4}[\delta\sqrt{V_1V_2} + \sqrt{V_1}w_{im} + \sqrt{V_1}w_m + \\
&\frac{1}{2}\delta^3\rho V_2 + \frac{3}{2}\rho(1+m)w_{im}^2 + \frac{3}{2}\delta\rho w_m^2 - 2\mu w_m^2] + \\
&\frac{1}{\epsilon H_{min}^5}[\mu(1+m)w_{im}^2 + \mu w_m^2] + \frac{1}{2\epsilon H_{min}^6}\delta\rho V_2 \\
&\leq O(\frac{\delta}{\epsilon H_{min}^2}) + O(\frac{1}{\epsilon H_{min}^3}) + O(\frac{\delta^3}{\epsilon H_{min}^4}) + \\
&O(\frac{1}{\epsilon H_{min}^5}) + O(\frac{\delta}{\epsilon H_{min}^6}).
\end{aligned}
\tag{23}
$$

Here we use O asymptotically upper bound: If there is a constant n_0, such that when $n \geq n_0$, there are $f(n) > 0$, $g(n) > 0$, and $\lim_{n\to\infty}\frac{f(n)}{g(n)} = 0$, we have $f(n) = O(g(n))$.

Therefore, QuAsyncFL converges to a critical point:

$$
\begin{aligned}
&\frac{1}{\sum_{k=0}^{K}H_k}\sum_{k=1}^{K}E[\|\nabla F(\mathbf{w}(t))\|_2^2] \\
&\leq O(\frac{\delta}{\epsilon H_{min}^2}) + O(\frac{1}{\epsilon H_{min}^3}) + O(\frac{\delta^3}{\epsilon H_{min}^4}) + \\
&O(\frac{1}{\epsilon H_{min}^5}) + O(\frac{\delta}{\epsilon H_{min}^6}).
\end{aligned}
\tag{24}
$$

This completes the proof.

References

[1] Z. Chang, S. Liu, X. Xiong, Z. Cai, and G. Tu, "A Survey of Recent Advances in Edge-Computing-Powered Artificial Intelligence of Things," *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 13 849–13 875, 2021.

[2] L. Jia, Z. Zhou, F. Xu, and H. Jin, "Cost-Efficient Continuous Edge Learning for Artificial Intelligence of Things," *IEEE Internet of Things Journal*, vol. 9, no. 10, pp. 7325–7337, 2022.

[3] M. A. Rahman, M. S. Hossain, A. J. Showail, N. A. Alrajeh, and A. Ghoneim, "AI-Enabled IIoT for Live Smart City Event Monitoring," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 2872–2880, 2023.

[4] Y. Liu, X. Ma, L. Shu, G. P. Hancke, and A. M. Abu-Mahfouz, "From Industry 4.0 to Agriculture 4.0: Current Status, Enabling Technologies, and Research Challenges," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4322–4334, 2021.

[5] X. Huang, Z. Hu, Y. Qiao, and S. Sukkarieh, "Deep Learning-Based Cow Tail Detection and Tracking for Precision Livestock Farming," *IEEE/ASME Transactions on Mechatronics*, pp. 1–9, 2022.

[6] J. Zhang and D. Tao, "Empowering Things With Intelligence: A Survey of the Progress, Challenges, and Opportunities in Artificial Intelligence of Things," *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 7789–7817, 2021.

[7] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. Vincent Poor, "Federated Learning for Internet of Things: A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021.

[8] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3347–3366, 2023.

[9] Z. Yan, D. Li, Z. Zhang, and J. He, "Accuracy-security tradeoff with balanced aggregation and artificial noise for wireless federated learning," *IEEE Internet of Things Journal*, pp. 1–1, 2023.

[10] X. Zhang, A. Mavromatics, A. Vafeas, R. Nejabati, and D. Simeonidou, "Federated Feature Selection for Horizontal Federated Learning in IoT Networks," *IEEE Internet of Things Journal*, pp. 1–1, 2023.

[11] H. Zhu, R. Wang, Y. Jin, and K. Liang, "PIVODL: Privacy-preserving vertical federated learning over distributed labels," *IEEE Transactions on Artificial Intelligence*, pp. 1–1, 2021.

[12] K. I.-K. Wang, X. Zhou, W. Liang, Z. Yan, and J. She, "Federated Transfer Learning Based Cross-Domain Prediction for Smart Manufacturing," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 6, pp. 4088–4096, 2022.

[13] C. You, D. Feng, K. Guo, H. H. Yang, C. Feng, and T. Q. S. Quek, "Semi-Synchronous Personalized Federated Learning Over Mobile Edge Networks," *IEEE Transactions on Wireless Communications*, vol. 22, no. 4, pp. 2262–2277, 2023.

[14] N. Yang, D. Yuan, Y. Zhang, Y. Deng, and W. Bao, "Asynchronous Semi-Supervised Federated Learning with Provable Convergence in Edge Computing," *IEEE Network*, vol. 36, no. 5, pp. 136–143, 2022.

[15] Y. Liu, L. Shu, Z. Huo, K. F. Tsang, and G. P. Hancke, "Collaborative Industrial Internet of Things for Noise Mapping: Prospects and Research Opportunities," *IEEE Industrial Electronics Magazine*, vol. 15, no. 2, pp. 52–64, 2021.

[16] Y. Liu, D. Li, B. Du, L. Shu, and G. Han, "Rethinking Sustainable Sensing in Agricultural Internet of Things: From Power Supply Perspective," *IEEE Wireless Communications*, vol. 29, no. 4, pp. 102–109, 2022.

[17] Y. Liu, D. Li, H. Dai, C. Li, and R. Zhang, "Understanding the Impact of Environmental Conditions on Zero-Power Internet of Things: An Experimental Evaluation," *IEEE Wireless Communications*, pp. 1–8, 2022.

[18] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated Learning in Mobile Edge Networks: A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 22, no. 3, pp. 2031–2063, 2020.

[19] C. Xie, S. Koyejo, and I. Gupta, "Asynchronous Federated Optimization," *CoRR*, vol. abs/1903.03934, 2019. [Online]. Available: http://arxiv.org/abs/1903.03934

[20] Y. Chen, Y. Ning, M. Slawski, and H. Rangwala, "Asynchronous Online Federated Learning for Edge Devices with Non-IID Data," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 15–24.

[21] S. Horvóth, C.-Y. Ho, L. Horvath, A. N. Sahu, M. Canini, and P. Richtárik, "Natural compression for distributed deep learning," in *Mathematical and Scientific Machine Learning*. PMLR, 2022, pp. 129–141.

[22] D. Zeng, S. Liang, X. Hu, H. Wang, and Z. Xu, "FedLab: A Flexible Federated Learning Framework," *Journal of Machine Learning Research*, vol. 24, no. 100, pp. 1–7, 2023.

[23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[24] Z. Wang, Z. Zhou, H. Zhang, G. Zhang, H. Ding, and A. Farouk, "AI-Based Cloud-Edge-Device Collaboration in 6G Space-Air-Ground Integrated Power IoT," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 16–23, 2022.

[25] Z. Ni, H. Chen, Z. Li, X. Wang, N. Yan, W. Liu, and F. Xia, "MSCET: A Multi-Scenario Offloading Schedule for Biomedical Data Processing and Analysis in Cloud-Edge-Terminal Collaborative Vehicular Networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2021.

[26] C. Qiu, X. Wang, H. Yao, J. Du, F. R. Yu, and S. Guo, "Networking Integrated Cloud–Edge–End in IoT: A Blockchain-Assisted Collective Q-Learning Approach," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12 694–12 704, 2021.

[27] Z. Zhong, W. Bao, J. Wang, X. Zhu, and X. Zhang, "FLEE: A Hierarchical Federated Learning Framework for Distributed Deep Neural Network over Cloud, Edge, and End Device," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 5, Oct 2022.

[28] Z. Ma, H. Li, W. Fang, Q. Liu, B. Zhou, and Z. Bu, "A Cloud-Edge-Terminal Collaborative System for Temperature Measurement in COVID-19 Prevention," in *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2021, pp. 1–6.

[29] Y. Liu, Z. Wang, X. Wu, F. Fang, and A. S. Saqlain, "Cloud-Edge-End Cooperative Detection of Wind Turbine Blade Surface Damage Based on Lightweight Deep Learning Network," *IEEE Internet Computing*, vol. 27, no. 1, pp. 43–51, 2023.

[30] J. Yang, Z. Guo, J. Luo, Y. Shen, and K. Yu, "Cloud-Edge-End Collaborative Caching Based on Graph Learning for Cyber-Physical Virtual Reality," *IEEE Systems Journal*, pp. 1–12, 2023.

[31] S. Zhang, Z. Wang, Z. Zhou, Y. Wang, H. Zhang, G. Zhang, H. Ding, S. Mumtaz, and M. Guizani, "Blockchain and Federated Deep Reinforcement Learning Based Secure Cloud-Edge-End Collaboration in Power IoT," *IEEE Wireless Communications*, vol. 29, no. 2, pp. 84–91, 2022.

[32] H. Liao, Z. Zhou, N. Liu, Y. Zhang, G. Xu, Z. Wang, and S. Mumtaz, "Cloud-Edge-Device Collaborative Reliable and Communication-Efficient Digital Twin for Low-Carbon Electrical Equipment Management," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1715–1724, 2023.

[33] Z. Zhou, Y. Li, X. Ren, and S. Yang, "Towards Efficient and Stable K-Asynchronous Federated Learning With Unbounded Stale Gradients on Non-IID Data," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 12, pp. 3291–3305, 2022.

[34] Q. Ma, Y. Xu, H. Xu, Z. Jiang, L. Huang, and H. Huang, "FedSA: A Semi-Asynchronous Federated Learning Mechanism in Heterogeneous Edge Computing," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3654–3672, 2021.

[35] T. Zhang, A. Song, X. Dong, Y. Shen, and J. Ma, "Privacy-Preserving Asynchronous Grouped Federated Learning for IoT," *IEEE Internet of Things Journal*, vol. 9, no. 7, pp. 5511–5523, 2022.

[36] C. Pan, Z. Wang, H. Liao, Z. Zhou, X. Wang, M. Tariq, and S. Al-Otaibi, "Asynchronous Federated Deep Reinforcement Learning-Based URLLC-Aware Computation Offloading in Space-Assisted Vehicular Networks," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2022.

[37] Q. Liu, B. Yang, Z. Wang, D. Zhu, X. Wang, K. Ma, and X. Guan, "Asynchronous Decentralized Federated Learning for Collaborative Fault Diagnosis of PV Stations," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 3, pp. 1680–1696, 2022.

[38] J. Lu, H. Liu, Z. Zhang, J. Wang, S. K. Goudos, and S. Wan, "Toward Fairness-Aware Time-Sensitive Asynchronous Federated Learning for Critical Energy Infrastructure," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3462–3472, 2022.

[39] J. Liu, H. Xu, L. Wang, Y. Xu, C. Qian, J. Huang, and H. Huang, "Adaptive Asynchronous Federated Learning in Resource-Constrained Edge Computing," *IEEE Transactions on Mobile Computing*, vol. 22, no. 2, pp. 674–690, 2023.

[40] C.-H. Hu, Z. Chen, and E. G. Larsson, "Scheduling and Aggregation Design for Asynchronous Federated Learning Over Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 4, pp. 874–886, 2023.

[41] H. Liu, F. He, and G. Cao, "Communication-Efficient Federated Learning for Heterogeneous Edge Devices Based on Adaptive Gradient Quantization," *arXiv preprint arXiv:2212.08272*, 2022.

[42] Z. Yan, D. Li, X. Yu, and Z. Zhang, "Latency-Efficient Wireless Federated Learning With Quantization and Scheduling," *IEEE Communications Letters*, vol. 26, no. 11, pp. 2621–2625, 2022.

[43] S. Jang and H. Lim, "AsyncFL: Asynchronous Federated Learning Using Majority Voting with Quantized Model Updates (Student Abstract)," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, pp. 12 975–12 976, Jun. 2022.

[44] J. Sun, T. Chen, G. B. Giannakis, Q. Yang, and Z. Yang, "Lazily Aggregated Quantized Gradient Innovation for Communication-Efficient Federated Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2031–2044, 2022.
[45] S. Chen, C. Shen, L. Zhang, and Y. Tang, "Dynamic Aggregation for Heterogeneous Quantization in Federated Learning," *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6804–6819, 2021.
[46] O. A. Hanna, Y. H. Ezzeldin, C. Fragouli, and S. Diggavi, "Quantization of Distributed Data for Learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 3, pp. 987–1001, 2021.
[47] Y. Du, S. Yang, and K. Huang, "High-Dimensional Stochastic Gradient Quantization for Communication-Efficient Edge Learning," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2128–2142, 2020.
[48] C. Dupuy, R. Arava, R. Gupta, and A. Rumshisky, "An Efficient DP-SGD Mechanism for Large Scale NLU Models," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4118–4122.
[49] Q. Lu and B. Murmann, "Enhancing the Energy Efficiency and Robustness of TinyML Computer Vision Using Coarsely-Quantized Log-Gradient Input Images," *ACM Transactions on Embedded Computing Systems*, Apr 2023.
[50] Z. Tang, S. Shi, X. Chu, W. Wang, and B. Li, "Communication-Efficient Distributed Deep Learning: A Comprehensive Survey," *arXiv preprint arXiv:2003.06307*, 2020.

Ye Liu received M.S. and Ph.D. degrees in electronic science and engineering from Southeast University, Nanjing, Nanjing, China, in 2013 and 2018, respectively. He was a Visiting Scholar with Montana State University, Bozeman, MT, the USA, from October 2014 to October 2015. He was a visiting Ph.D. Student from February 2017 to January 2018 with the Networked Embedded Systems Group, RISE Swedish Institute of Computer Science. He has authored or co-authored papers in several prestigious journals and conferences, such as the IEEE WCM, IEEE IEM, IEEE ComMag, IEEE NetMag, IEEE TOM, IEEE IoTJ, ACM TECS, ACM TOSN, INFOCOM, IPSN, ICNP, and EWSN. His current research interests include wireless sensor networks, energy harvesting systems, and mobile crowdsensing. Dr. Liu was awarded first place in the EWSN Dependability Competition in 2019 and the Macao Young Scholar in 2021.

Peishan Huang received her M.Sc. degree from the Macau University of Science and Technology in 2022. Currently, she is pursuing her Ph.D. degree at the Macau University of Science and Technology. Her research interests include federated learning, wireless communications, and machine learning.

Fan Yang received the B.S. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2011 , the M.S. and Ph.D. degrees from Nanjing Agricultural University, Nanjing, China, in 2014 and 2022, respectively. He is currently a lecturer with School of Mathematics and Statistics, Jiangsu Normal University, Xuzhou, China. His research interests include performance analysis and protocol design on networks, Wireless Sensor Networks and IoT.

Kai Huang received the Ph.D. degree in agricultural engineering from China Agricultural University, Beijing, China, in 2018. He is currently a researcher at Institute of Agricultural Facilities and Equipment, Jiangsu Academy of Agricultural Sciences, Nanjing, China. His main research fields include Agricultural Internet of Things, Smart Phytoprotection. He has guest edited the special issue "Smart Agricultural Applications with Internet of Things" in Sensors. He has served as a PC Member in 3PGCIC 2019, C4W 2020, and a TPC Member in ICIN 2021, ICC 2021. He has served as a ITIA 2022 TPC Co-Chair. He has served as a reviewer editor in the journal of Frontiers in Plant Science.

Lei Shu received the B.S. degree in computer science from South Central University for Nationalities, China, in 2002, the M.S. degree in computer engineering from Kyung Hee University, South Korea, in 2005, and the Ph.D. degree from the Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland, in 2010. Until 2012, he was a Specially Assigned Researcher with the Department of Multimedia Engineering, Graduate School of Information Science and Technology, Osaka University, Japan. He is currently a Distinguished Professor with Nanjing Agricultural University, China, and a Lincoln Professor with the University of Lincoln, U.K. His current research interests include wireless sensor networks and the Internet of Things.