This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCCN.2025.3586904, IEEE Transactions on Cognitive Communications and Networking

Personalized Federated Learning for GAI-Assisted Semantic Communications

Yubo Peng, *Student Member, IEEE*, Feibo Jiang, *Senior Member, IEEE*, Li Dong, Kezhi Wang, *Senior Member, IEEE*, and Kun Yang, *Fellow, IEEE*.

***Abstract*—Semantic Communication (SC), which focuses on transmitting meaning rather than raw data, has emerged as the next-generation communication paradigm. However, the performance of SC is heavily influenced by network design and training methodologies. To address these challenges and enhance SC performance at the edge, we first introduce a Generative Artificial Intelligence (GAI)-assisted SC (GSC) model, which improves SC capabilities by optimizing the network architecture. Then, to achieve the efficient learning of GSC models deployed on each user, a Personalized Semantic Federated Learning (PSFL) framework is proposed. Specifically, in the local training phase, a Personalized Local Distillation (PLD) approach is employed, where each user selects a tailored GSC model as a mentor based on local resources. This mentor subsequently distills knowledge to a unified student model, ensuring compliance with the model isomorphism requirements of FL. In the global aggregation phase, an Adaptive Global Pruning (AGP) scheme is applied, dynamically pruning or expanding the aggregated global model based on real-time channel conditions. This mechanism effectively balances accuracy and communication energy efficiency. Finally, numerical results validate the feasibility and efficacy of the proposed PSFL framework, demonstrating its potential to enhance SC performance in edge environments significantly.**

***Index Terms*—Semantic communication; federated learning; generative artificial intelligence; network pruning.**

I. Introduction

As an innovative communication paradigm in 6G, Semantic Communication (SC) becomes one of the intelligent solutions for the spectrum sacrifice caused by the various emerging applications on users [1]. Different from traditional communication, SC aims to transmit only semantically related information for the respective task/goal [2]. For example, in the fault detection scenario, the users first extract semantic information from the surveillance video by the deployed SC encoder (i.e., semantic and channel encoders), then just transmit slightly semantic information to the Edge Server (ES) deployed on the Base Station (BS). Finally, the received semantic information is decoded by the SC decoder (i.e., semantic and channel decoders) deployed on the ES. Since the transferred data is greatly reduced in SC, the consumption of spectrum resources is also correspondingly greatly decreased.

The performance of SC is highly dependent on the construction of high-quality SC models, hence many researchers construct SC models based on Deep Learning (DL) models. For instance, Xie *et al*. [3] proposed a DL-based SC system, aiming to maximize system capacity and minimize semantic errors in text transmission by restoring the meaning of sentences. Wang *et al*. [4] optimized DL-based joint source-channel coding by introducing adversarial loss, which better preserved the global semantic information and local texture details of images. Han *et al*. [5] proposed a novel end-to-end DL-based speech-oriented SC system, utilizing a soft alignment module and a redundancy removal module to extract text-related semantic features while discarding semantically redundant content. Most of the above works are based on traditional discriminative Artificial Intelligence (AI) methods, which typically involve small models trained for specific application scenarios. This approach inherently limits their adaptability across different environments. Moreover, discriminative AI primarily focuses on learning local and short-term features, leading to challenges such as getting trapped in local minima and exhibiting limited generative capabilities [6], [7].

Generative AI (GAI), as a recent advancement in AI technology, not only possesses remarkable generative capabilities but also exhibits more powerful data processing abilities than discriminative AI [8]. The latest GAI models, such as GPT-4 and LLaMA 3.1 [9], have been widely applied across various domains. Therefore, it has become a recent research topic that construct SC models based on GAI. Du *et al*. [10] designed an AI-generated incentive mechanism based on the diffusion model in full-duplex end-to-end SC to promote semantic information sharing among users. Lin *et al*. [11] proposed a blockchain-assisted SC framework for AI-generated content services to address security issues arising from malicious semantic data transmission in SC. Guo *et al*. [12] introduced a semantic importance-aware communication scheme using pre-trained language models to quantify the semantic importance of data frames, thereby reducing semantic loss in communication.

In addition to the network structure, the training method is also a key factor affecting the performance of SC [13].

This paper was partly funded by Jiangsu Major Project on Basic Research (Grant No.: BK20243059), Gusu Innovation Project for People (Grant No.: ZXL2024360), Natural Science Foundation of China (Grant No. 62132004), the Major Program Project of Xiangjiang Laboratory (Grant No. XJ2023001 and XJ2022001), and Qiyuan Lab Innovation Fund (Grant No. 2022-JCJQ-LA-001-088). (Corresponding author: Feibo Jiang.)

Yubo Peng (ybpeng@smail.nju.edu.cn) and Kun Yang (kunyang@nju.edu.cn) are with the State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, 210008, China, and School of Intelligent Software and Engineering, Nanjing University (Suzhou Campus), Suzhou, 215163, China.

Feibo Jiang (jiangfb@hunnu.edu.cn) is with the School of Information Science and Engineering, Hunan Normal University, Changsha, China.

Li Dong (Dlj2017@hunnu.edu.cn) is with Changsha Social Laboratory of Artificial Intelligence, Hunan University of Technology and Business, Changsha, China.

Kezhi Wang (Kezhi.Wang@brunel.ac.uk) is with the Department of Computer Science, Brunel University London, UK.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCCN.2025.3586904, IEEE Transactions on Cognitive Communications and Networking

However, the traditional centralized learning method requires the users to transmit the local data to the central server for centralized training, which may lead to a high consumption of communication energy and a high risk of information leakage [14]. Hence, this traditional approach is not suitable for users to train the SC models at the edge.

Federated Learning (FL) [15] has the potential to alleviate the above issues. FL enables several clients and a central server to train the SC models collaboratively only by sharing model parameters, rather than transmitting large raw training data. Numerous studies have focused on communication-efficient FL. For example, Nguyen *et al*. [16] presented a high-compression FL scheme that effectively reduced data load during FL processes without modifying the structure or hyperparameters. Similarly, Wang *et al*. [17] proposed a communication-efficient adaptive federated optimization method that substantially lowered communication costs through error feedback and compression strategies. Furthermore, Hönig *et al*. [18] developed a doubly-adaptive quantization FL algorithm that dynamically adjusted the quantization level over time and among various clients, enhancing compression while maintaining model quality. Although these studies present efficient FL algorithms to enhance model training, they ignore the issues of model adaptation for heterogeneous users and the high communication overheads in dynamic networks.

Based on the above review of related work, we summarize three critical challenges that apply SC to users as follows:

1) *Insufficient semantic extraction capabilities:* Traditional discriminative network architectures face significant limitations in effective semantic extraction, particularly in complex communication scenarios. For instance, while Convolutional Neural Networks (CNNs) excel at capturing local features due to their hierarchical structure, they struggle to effectively capture global contextual information [19]. This inherent limitation hampers their ability to construct comprehensive semantic representations of input data.
2) *Model adaptation for heterogeneous devices in FL:* Users are usually heterogeneous, which means they have different scales of local data and computation resources. Generally, more complex models can achieve higher accuracy when the data and computation resources are enough [20]. Hence, the users, having more available data and computation resources, may need sophisticated models to achieve higher accuracy. However, the limited-resource users can only use a compact model for local training, and the well-resourced users have to choose the same compact model as a compromise to meet the model isomorphism requirement of FL.
3) *High communication overheads in dynamic networks:* While conventional FL algorithms facilitate distributed training using local data from multiple users, ensuring data privacy and security, they often result in substantial network traffic and communication overhead due to frequent parameter exchanges [21]. Communication-efficient FL methods, such as those in [17] and [18], alleviate communication energy consumption by compressing transmitted parameters. However, these compression techniques are performed on the client side, leading to additional costs for clients. Moreover, these methods do not account for the impact of fluctuating network conditions, such as variations in Signal-to-Noise Ratio (SNR) in wireless communications.

To solve the above issues, we first design a novel GAI-assisted SC (GSC) model to apply in the communications between users and BS, improving the utilization of limited spectrum resources. Then, a Personalized Semantic Federated Learning (PSFL) is proposed, where improved local training and global aggregation methods are employed to train the GSC models deployed on users while protecting privacy and security. The main contributions are summarized as follows:

1) *Accurate semantic transmission*: Considering the shortcomings of the discriminative network, we employed GAI networks in both the semantic encoder and decoder in the GSC model. Specifically, we employ the Vision Transformer (ViT), a common GAI network, for processing images. This can achieve more accurate semantic feature extraction of transmitted images at the transmitter and more precise image reconstruction at the receiver through the multi-head self-attention mechanism. Therefore, the proposed first challenge is solved.
2) *High-quality local training*: We propose a Personalized Local Distillation (PLD) strategy in the local training phase of PSFL, improving the accuracy of the GSC model. In PLD, each user can select a suitable GSC model as a mentor according to their local resources and a unified CNN-based SC (CSC) model as a student. Then, the mentor model can be distilled to the student model to meet the model isomorphism requirement of the FL. As a result, PLD addresses the second challenge.
3) *Energy-efficent global aggregation*: We design an Adaptive Global Pruning (AGP) algorithm in the global aggregation phase of PSFL, reducing the consumption of communication energy. Specifically, the aggregated global FL model (i.e., the updated CSC model) is pruned or expanded. The pruning ratio is determined by considering the real-time SNR between users and the BS. Therefore, the AGP solves the last challenge.

The rest of this paper can be organized as follows. The system model is introduced in Section II. The proposed PSFL is described in Section III. The complexity analysis is introduced in Section IV. Numerical results are presented in Section V. The work summary and future expectations are described in Section VI.

II. System Model

Fig. 1 illustrates the communication between users and the BS through the SC system. We consider an uplink wireless network with limited spectrum resources to deploy a distributed SC system, comprising K users, denoted by the set $\mathcal{K}$, and a single BS with an ES. In the training phase, as shown in Fig. 1(a), the BS is responsible for performing global aggregation and updating the global SC model, while the users train their local SC models based on local data and subsequently transmit

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCCN.2025.3586904, IEEE Transactions on Cognitive Communications and Networking

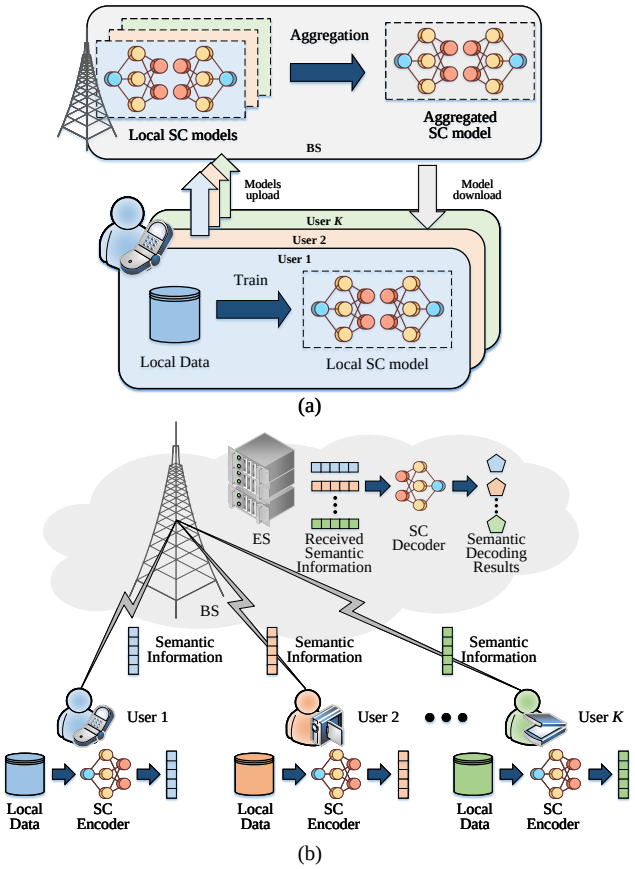


Fig. 1: The illustration that users communicate with BS using SC. (a) Training phase. (b) Inference phase.

the model parameters to the BS. In the inference phase, as shown in Fig. 1(b), the users only need to transmit semantic information to the BS during data transmission, rather than the large-sized raw data [22]. This semantic information is then decoded at the ES. To facilitate the extraction of semantic information, the SC encoder is deployed on each user, while the SC decoder is deployed on the ES to decode the received semantic information. Additionally, we consider the impairments of the physical channels between the users and the BS.

A. GSC Model

We mainly consider the image SC that refers to capturing the semantics of interest in input images, thereby reducing the amount of data required for image transmission and conserving bandwidth. Compared to traditional CNNs, ViT displays superior feature analysis capabilities in various visual tasks, such as image classification, object detection, and feature extraction [23]. Therefore, as shown in Fig. 2, the GSC model employs ViT as the image semantic encoder and decoder. Subsequently, Deep Neural Networks (DNNs) construct the channel encoder and decoder. Finally, a perception model simulates the physical channel, ensuring it supports backpropagation. The transmission process of image SC based on the GSC model is as follows:

1) Transmitter: First, the input image $\mathbf{m} \in \mathbb{R}^{H\times W\times C}$ is inputted into a PatchEmbed layer, in which $\mathbf{m}$ is converted into N patches of size $(P^2 \times C)$, where H denotes the height of the image, W denotes the width, C denotes the number of channels, and P^2 represents the number of segments into which the image is divided. Hence, $N = \frac{W\times H}{P^2}$.

Then, the sequence X_p composed of these N patches undergoes the Patch Embedding operation [24]. Specifically, each patch in X_p bypasses a linear transformation, reducing the dimensionality of the sequence to D and resulting in a linear embedding sequence:

$$\mathbf{Z}_0 = \left[X_p^1 E; X_p^2 E; \ldots; X_p^N E\right], \tag{1}$$

where E is the linear transformation (i.e., a fully connected layer) with input dimensions $(P^2 \times C)$ and output dimensions N. X_p^i represents the i-th patch in X_p.

Next, the position vectors from the positional encoding, which contain positional information, are linearly combined with $\mathbf{Z}_0$ to obtain the input sequence of ViT as follows:

$$\mathbf{Z} = \left[X_p^1 E + P_1; X_p^2 E + P_2; \ldots; X_p^N E + P_N\right], \tag{2}$$

where P_i denotes the i-th position vector. Subsequently, the semantic encoder extracts features from $\mathbf{Z}$, resulting in the semantic feature of the original image $\mathbf{m}$. The image semantic encoder is based on ViT, with its core being the multi-head attention layer, which can learn the relationships between pixels through the multi-head attention mechanism, enabling more accurate feature extraction and realistic image reconstruction. The multi-head attention layer is essentially composed of multiple self-attention heads concatenated together. A self-attention head is derived from a single self-attention layer, which can be calculated by:

$$head_i = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{3}$$

where $\mathbf{Q}$ is the query vector, $\mathbf{K}$ is the matching vector corresponding to $\mathbf{Q}$, and $\mathbf{V}$ is the information vector. $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are all obtained through linear transformations of the input $\mathbf{Z}$, i.e., $\mathbf{Q} = \mathbf{Z}\mathbf{W}_\mathbf{Q}$, $\mathbf{K} = \mathbf{Z}\mathbf{W}_\mathbf{K}$, and $\mathbf{V} = \mathbf{Z}\mathbf{W}_\mathbf{V}$, where $\mathbf{W}_\mathbf{Q}$, $\mathbf{W}_\mathbf{K}$, and $\mathbf{W}_\mathbf{V}$ are the respective weight matrices. d_k is the scaling factor. Assuming the number of heads is h, the multi-head self-attention can be calculated by:

$$\begin{aligned}&\text{MultiheadAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \\ &\text{concat}(head_1, \cdots, head_h)\mathbf{W}_{\text{mha}},\end{aligned} \tag{4}$$

where $\mathbf{W}_{\text{mha}}$ represents the weights of the multi-head attention layer.

Finally, to ensure data is transmitted on the physical channel, the semantic feature should be encoded and modulated by the channel encoder to reduce channel impairments and improve robustness. The transmitted signal can be represented as:

$$\mathbf{X} = C\left(S\left(\mathbf{Z}, \boldsymbol{\vartheta}\right), \boldsymbol{\alpha}\right), \tag{5}$$

where $S(\cdot)$ represents the semantic encoder with model parameters $\boldsymbol{\vartheta}$ and $C(\cdot)$ is the channel encoder with model parameters $\boldsymbol{\alpha}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCCN.2025.3586904, IEEE Transactions on Cognitive Communications and Networking

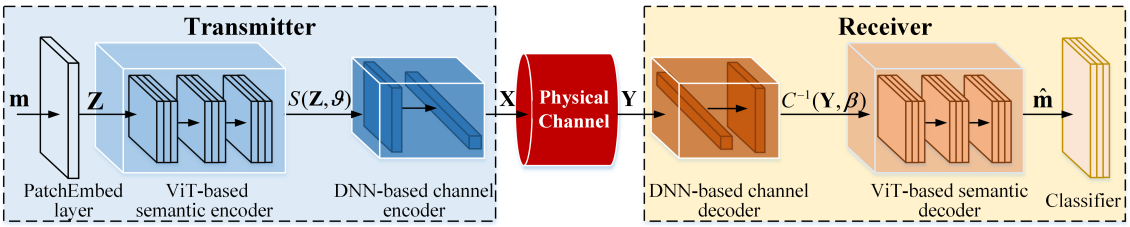


Fig. 2: The illustration of image transmission utilizing the proposed GSC model.

2) Physical channel: When transmitted over the physical channel, $\mathbf{X}$ suffers transmission impairments that include attenuation and noise. The transmission process of the physical channel can be expressed as:

$$\mathbf{Y} = \mathbf{H} \cdot \mathbf{X} + \mathbf{N}, \tag{6}$$

where $\mathbf{Y}$ represents the received signal; $\mathbf{H}$ represents the channel gain between the transmitter and the receiver; $\mathbf{N}$ is Additive White Gaussian Noise (AWGN). For end-to-end training of encoder and decoder, the physical channel must allow back-propagation [3], hence we use a perception model to simulate the physical channel.

3) Receiver: The channel decoder demodulates the received signal to extract the semantic features, which are then decoded by the semantic decoder. The recovered image, $\hat{\mathbf{m}}$, is obtained as follows:

$$\hat{\mathbf{m}} = S^{-1}\left(C^{-1}(\mathbf{Y}, \boldsymbol{\beta}), \boldsymbol{\delta}\right), \tag{7}$$

where $C^{-1}(\cdot)$ represents the channel decoder with model parameters $\boldsymbol{\beta}$; $S^{-1}(\cdot)$ is the semantic decoder with model parameters $\boldsymbol{\delta}$.

B. FL Model

The local dataset of the k-th user defines as $\mathcal{D}_k = \{(\mathbf{m}_{k,1}, y_{k,1}), (\mathbf{m}_{k,2}, y_{k,2}), ..., (\mathbf{m}_{k,N_k}, y_{k,N_k})\}$, where N_k is the number of collected samples in $\mathcal{D}_k$, $\mathbf{m}_{k,i}$ is the i-th sample and $y_{k,i}$ is the corresponding label. Note that $\mathcal{D}_k$ may be non-Independent Identical Distributed (non-IID) data, which depends on the realistic environment and usage pattern of the k-th user.

For the k-th user, the local loss function in the t-th communication round of FL can be calculated as:

$$F_k\left(\mathbf{w}_{k,t}\right) = \frac{1}{N_k} \sum_{i=1}^{N_k} f\left(\mathbf{w}_{k,t}, \mathbf{m}_{k,i}, y_{k,i}\right), \tag{8}$$

where $f(\mathbf{w}_{k,t}, \mathbf{m}_{k,i}, y_{k,i})$ is the training loss function for the i-th sample $(\mathbf{m}_{k,i}, y_{k,i})$ in $\mathcal{D}_k$; $\mathbf{w}_{k,t}$ is the weights of the local FL model of the k-th user in the t-th communication round. $\mathbf{w}_{k,t}$ includes all the parameters of the GSC model, namely $\mathbf{w}_{k,t} = (\boldsymbol{\alpha}_{k,t}, \boldsymbol{\vartheta}_{k,t}, \boldsymbol{\beta}_{k,t}, \boldsymbol{\delta}_{k,t})$, where $\boldsymbol{\alpha}_{k,t}, \boldsymbol{\vartheta}_{k,t}, \boldsymbol{\beta}_{k,t}, \boldsymbol{\delta}_{k,t}$ represent the parameters of the GSC model deployed on the k-th user from the channel encoder to the semantic decoder in the t-th communication round.

In this system, we consider the classification tasks-oriented SC. Specifically, we use a pre-trained classifier (e.g., ResNet-18 [25]) to perform image classification based on $\hat{\mathbf{m}}$ and the cross-entropy (CE) as the loss function of the GSC model, hence $F_k\left(\mathbf{w}_{k,t}\right)$ can be calculated by:

$$F_k\left(\mathbf{w}_{k,t}\right) = \mathcal{L}_{\mathrm{CE}}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{i=1}^{M} y_i \log\left(\hat{y}_i\right), \tag{9}$$

where $\mathbf{y} = [y_1, y_2, ..., y_M]$ represents the represents the predicted probabilities based on the raw image $\mathbf{m}$, when the input data m belongs to the i-th class, $y_i = 1$, otherwise 0; $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_M]$ represents the predicted probabilities based on the reconstructed image $\hat{\mathbf{m}}$, $\hat{y}_i$ represents the probability predicted as the i-th class; M is the total number of categories.

To ensure data privacy and security, the global FL model is updated in ES by aggregating the local FL models from users. The update of the global FL model can be given by [15]:

$$\mathbf{w}_{g,t} = \frac{\sum_{k=1}^{K} N_k \mathbf{w}_{k,t}}{\sum_{k=1}^{K} N_k}, \tag{10}$$

where $\mathbf{w}_{g,t}$ is the weights of the global FL model in the t-th communication round. Note that $\mathbf{w}_{g,t}$ shares the same architecture to $\mathbf{w}_{k,t}$. In addition, since the transmitted parameters in FL may also leak sensitive data, as a solution, the differential privacy and homomorphic encryption algorithms can be used to encrypt the transmitted parameters and improve the data security and privacy [26]. Specifically, we can adopt an additive homomorphic encryption scheme (e.g., Paillier encryption) to encrypt local model updates $\mathbf{w}_{k,t}$ before transmission. This allows the edge server (ES) to perform secure aggregation directly on the encrypted parameters without decrypting them, thus preventing information leakage during communication. Furthermore, to prevent inference attacks from the aggregated model, we apply (ε, δ)-differential privacy to the local updates by adding calibrated Gaussian noise to the gradients before encryption. This combination ensures that both the individual contributions of users and the transmitted parameters are protected, thereby enhancing the overall privacy guarantee of the federated learning process.

FL aims to get the optimal global model $\mathbf{w}_{g,t}$ that can minimize the local FL loss of all devices, so as to achieve global optimization. Hence, the global loss function of FL can be given by:

$$F_g\left(\mathbf{w}_{g,t}\right) = \frac{1}{K} \sum_{k=1}^{K} F_k\left(\mathbf{w}_{g,t}\right), \tag{11}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCCN.2025.3586904, IEEE Transactions on Cognitive Communications and Networking

where $F_k(\mathbf{w}_{g,t})$ represents that the local FL loss based on $\mathbf{w}_{g,t}$. We take minimizing $F_g(\mathbf{w}_{g,t})$ as the goal of training the GSC model.

C. Communication Model

We consider that Orthogonal Frequency Division Multiple Access (OFDMA) is adopted for the links between users and the BS. When the k-th user uploads its local model weights $\mathbf{w}_{k,t}$ to BS, the uplink rate can be given by:

$$v_{k,t} = B_{k,t} \log_2(1 + \psi_{k,t}), \tag{12}$$

where $B_{k,t}$ and $\psi_{k,t}$ represent the uploading bandwidth and SNR of the k-th user in the t-th communication round, respectively. The transmission delay between the k-th user and the BS over uplink in the t-th communication round is calculated as:

$$\tau_{k,t} = \frac{Z(\mathbf{w}_{k,t})}{v_{k,t}}, \tag{13}$$

where $Z(\mathbf{w}_{k,t})$ represents the number of bits that each user k requires transmitting to the BS. The energy consumption of the communication process can be given by:

$$E_{k,t}^{\text{com}} = P_{k,t}\tau_{k,t}, \tag{14}$$

where $P_{k,t}$ is the transmitting power of the k-th user in the t-th communication round. The communication energy of FL is another critical optimization goal.

D. Problem Formulation

In consequence, our goal is to minimize the global loss function of FL and the communication energy consumption of the entire FL training process. The objective function can be defined as follows:

$$\min_{\mathbf{w}_{g,T}} F_g(\mathbf{w}_{g,T}) + \eta \sum_{t=1}^{T}\sum_{k=1}^{K} E_{k,t}^{\text{com}} \tag{15}$$

$$\text{s.t. } \tau_{k,t} \le \tau_k^{\text{req}}, \ \forall k \in \mathcal{K} \tag{16}$$

$$E_{k,t}^{\text{com}} \le E_{k,t}^{\text{rest}}, \ \forall k \in \mathcal{K} \tag{17}$$

$$0 \le P_{k,t} \le P_k^{\max}, \ \forall k \in \mathcal{K}, \tag{18}$$

where T represents the total number of communication rounds in FL; $\mathbf{w}_{g,T}$ represents the global FL model weights in the T-th communication round, $F_g(\mathbf{w}_{g,T})$ is the final global FL loss based on $\mathbf{w}_{g,T}$; η is a coefficient that adjusts the sensitivity of $E_{k,t}^{\text{com}}$ so that $F_g(\mathbf{w}_{g,T})$ and $E_{k,t}^{\text{com}}$ are relatively balanced; τ_k^{req} is the delay requirement of the k-th user for implementing the FL algorithm; $E_{k,t}^{\text{rest}}$ is the rest energy of the k-th user in the t-th communication round; $P_k^{\max}$ represents the maximum transmit power of the k-th user. Eq. (16) is the delay constraint of executing the FL in each communication round. Eq. (17) is the energy consumption constraint of performing FL in each communication round. Eq. (18) is the maximum transmit power constraint for users.

III. Personalized Semantic Federated Learning for GSC Model

To address the challenges of training GSC models deployed on users, we propose the PSFL, consisting of the PLD strategy and AGP algorithm, to optimize FL in the phases of local training and global aggregation, respectively.

A. PLD for Local Training

To address the issue of model adaptation for heterogeneous users and ensure effective information exchange among different GSC models, we propose the PLD strategy during the local training phase. Specifically, as shown in Fig. 3, each user not only deploys a suitable GSC model based on their local resources but also deploys a unified and small-scale CSC model. The GSC model is used for SC services after training. The CSC model, serving as a vehicle for transferring knowledge of the GSC model, is uploaded to the BS for parameter aggregation. It is then returned to the local user, transmitting the newly aggregated knowledge back to the GSC model. This process indirectly facilitates the information exchange of heterogeneous GSC models across different users. To achieve effective knowledge exchange between the GSC and CSC models in this process, Knowledge Distillation (KD) is utilized.

KD is a transfer learning method involving a sophisticated mentor model and a compact student model, aiming to transfer knowledge from the mentor to the student model [27]. In PLD, the GSC model acts as the mentor while the CSC model is the student. In PLD, the mutual learning process between the mentor and student models is as follows:

1) Distill knowledge from hard labels: The mentor and student models compute the loss between the output of the models and hard labels [28]. Generally, the hard labels are determined by the specific task. Since we consider the classification task, the hard labels are the categories of the input data. The input data on the k-th user is denoted as $\mathbf{m}_k$, and the corresponding hard labels are $\mathbf{y}_k$. The communication loss functions for classification tasks of mentor and student models are expressed as follows:

$$\mathcal{L}'_{\text{task}} = \mathcal{L}_{\text{CE}}(\mathbf{y}_k, \hat{\mathbf{y}}'_k) + \text{MSE}(\mathbf{m}_k, \hat{\mathbf{m}}'_k), \tag{19}$$

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{CE}}(\mathbf{y}_k, \hat{\mathbf{y}}_k) + \text{MSE}(\mathbf{m}_k, \hat{\mathbf{m}}_k), \tag{20}$$

where $\hat{\mathbf{y}}'_k$ and $\hat{\mathbf{y}}_k$ represent the probabilities of $\mathbf{m}_k$ as predicted by the mentor and student models, respectively. Similarly, $\hat{\mathbf{m}}'_k$ and $\hat{\mathbf{m}}_k$ denote the reconstructed images generated by the mentor and student models after wireless communications, respectively. The function $\text{MSE}(\cdot)$ represents the mean-square error, which is employed to ensure consistency between the original and reconstructed images at the pixel level. In summary, the task losses provide direct task-specific supervision for the mentor and student models.

2) Distill knowledge from soft labels: The mentor and student models transfer knowledge from the soft labels (e.g., $\hat{\mathbf{y}}'_k$ and $\hat{\mathbf{y}}_k$) reciprocally [28]. Since incorrect predictions from the mentor/student model may mislead the other one in the KD, we propose an adaptive method to weigh the distillation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCCN.2025.3586904, IEEE Transactions on Cognitive Communications and Networking

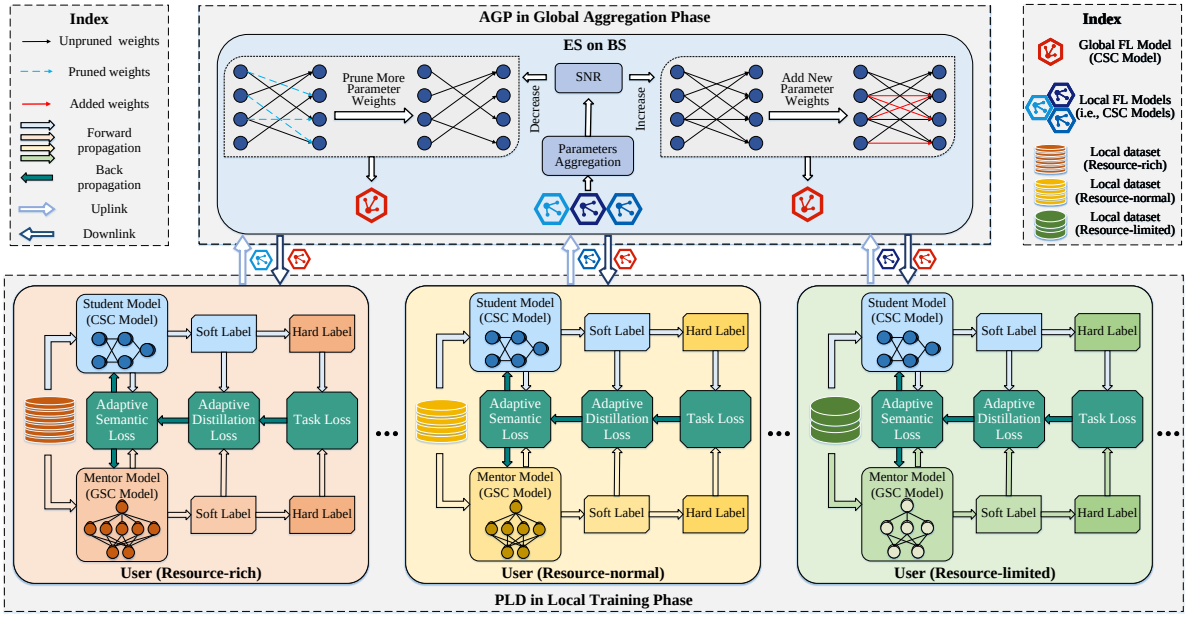


Fig. 3: The illustration of the proposed PSFL.

loss according to the quality of predicted soft labels (i.e., Eqs. (19) and (20)). The adaptive distillation losses of mentor and student models are formulated as follows:

$$\mathcal{L}'_{\text{dis}} = \frac{\text{KL}(\hat{\mathbf{y}}'_k, \hat{\mathbf{y}}_k)}{\mathcal{L}_{\text{task}}}, \tag{21}$$

$$\mathcal{L}_{\text{dis}} = \frac{\text{KL}(\hat{\mathbf{y}}_k, \hat{\mathbf{y}}'_k)}{\mathcal{L}'_{\text{task}}}, \tag{22}$$

where $\text{KL}(\cdot)$ means the Kullback–Leibler divergence. In this way, the distillation intensity is weak if the predictions of the mentor and student models are not reliable (e.g., their task losses are large). The distillation loss becomes dominant if the mentor and student models are well-tuned, which means small task losses. Thus, the adaptive distillation losses have the potential to mitigate the risk of overfitting.

3) Distill knowledge from the semantic information: To improve the performance of SC, the mentor and student models learn to minimize the difference between the output of the semantic encoder and the output of the channel decoder. Similarly, to avoid the misguiding of the mentor/student models in the interactive process, we also weigh the semantic loss according to task losses. Therefore, the adaptive semantic losses for the mentor and student models are formulated as follows:

$$\mathcal{L}'_{\text{sem}} = \mathcal{L}_{\text{sem}} = \frac{\text{MSE}(\mathbf{S}'_k, \mathbf{S}_k) + \text{MSE}(\mathbf{C}'_k, \mathbf{C}_k)}{\mathcal{L}'_{\text{task}} + \mathcal{L}_{\text{task}}}, \tag{23}$$

where $\mathbf{S}'_k$ and $\mathbf{S}_k$ represent the semantic encodings of the mentor and student models, respectively; $\mathbf{C}'_k$ and $\mathbf{C}_k$ represent the channel decodings of the mentor and student models, respectively.

4) Update mentor and student models: According to the above-proposed loss functions, the total loss functions for updating the mentor and student models are formulated as follows:

$$\mathcal{L}'_{\text{total}} = \mathcal{L}'_{\text{task}} + \mathcal{L}'_{\text{dis}} + \mathcal{L}'_{\text{sem}}, \tag{24}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{sem}}. \tag{25}$$

The mentor and student models update their weights by minimizing the total losses with the Stochastic Gradient Descent (SGD) optimizer [29]. We assume the training is performed on the k-th user with the dataset $\mathcal{D}_k$, the weights of the mentor and student models are denoted as $\mathbf{w}'_k$ and $\mathbf{w}_k$, respectively. G is used to denote the total epochs of the local training. The entire workflow of PLD is summarized in **Algorithm 1**.

Algorithm 1 PLD

Input: G, $\mathcal{D}_k$.
Output: $\mathbf{w}'_k$, $\mathbf{w}_k$.

for each epoch in G **do**
  for $i = 1, 2, ..., N_k$ **do**
    Sample $m_{k,i}$ from $\mathcal{D}_k$.
    Compute task losses $\mathcal{L}'_{\text{task}}$ and $\mathcal{L}_{\text{task}}$ according to Eqs. (19) and (20).
    Compute adaptive distillation losses $\mathcal{L}'_{\text{dis}}$ and $\mathcal{L}_{\text{dis}}$ according to Eqs. (21) and (22).
    Compute adaptive semantic losses $\mathcal{L}'_{\text{sem}}$ and $\mathcal{L}_{\text{sem}}$ according to Eq. (23).
    Compute total losses $\mathcal{L}'_{\text{total}}$ and $\mathcal{L}_{\text{total}}$ according to Eqs. (24) and (25).
    Update $\mathbf{w}_k$ by minimizing $\mathcal{L}_{\text{total}}$.
    Update $\mathbf{w}'_k$ by minimizing $\mathcal{L}'_{\text{total}}$.
  end for
end for

Overall, the Progressive Layer Dropping (PLD) strategy offers significant benefits in heterogeneous edge environments.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCCN.2025.3586904, IEEE Transactions on Cognitive Communications and Networking

By dynamically adjusting the number of active layers during local training based on the computational capabilities and energy constraints of each device, PLD effectively reduces the local computation overhead without relying on static pruning or model compression. This adaptive mechanism ensures that resource-constrained devices can still participate in federated learning without compromising model performance, thereby improving the overall system adaptability and efficiency.

B. AGP for Global Aggregation

In wireless environments, the traditional FL may bring unaffordable communication energy consumption for users due to the frequent parameter exchanges. From above Eqs. (12)-(14), the smaller $Z(\mathbf{w}_{k,t})$ and larger $\psi_{k,t}$ mean lower communication energy. Namely, when $\psi_{k,t}$ increases, more parameters can be transmitted; otherwise, fewer parameters, thereby ensuring low consumption of communication energy in dynamic SNR. Hence, the AGP algorithm is used to adaptively prune the global FL model $\mathbf{w}_{g,t}$ (i.e., the CSC model) after global aggregation. The proportion of pruning is determined by assessing the real-time SNR between users and the BS. As shown in Fig. 3, assuming that PSFL starts from a certain round t, the workflow of the PSFL assisted by AGP is described as follows:

1) Model pruning and weights broadcasting: We perform pruning on the global FL model $\mathbf{w}_{g,t}$ in the ES, in which a proportion of the smallest positive weights and the largest negative weights of $\mathbf{w}_{g,t}$ will be pruned. To avoid impairing the knowledge carried by the global FL model due to excessive pruning of parameters, the pruning proportion adaptively adjusts based on the real-time SNR between users and the BS in each communication round. Consequently, the pruning proportion in the t-th communication round can be represented as:

$$\zeta_t = \frac{\psi_{\max} - \frac{1}{K}\sum\limits_{k\in\mathcal{K}} \psi_{k,t}}{\psi_{\max} - \psi_{\min}}, \tag{26}$$

where $\psi_{\max}$ and $\psi_{\min}$ represent the maximum and minimum SNR between the users and BS, respectively. Thereafter, BS broadcasts the pruned global FL model $\hat{\mathbf{w}}_{g,t}$ to all users for the $(t+1)$-th round training.

2) Local training and weights uploading: Each user performs local training with the PLD strategy and obtains the latest mentor and student models. Since the FL model is pruned, it can be denoted as $\hat{\mathbf{w}}_{k,t+1}$. Afterward, the local FL model $\hat{\mathbf{w}}_{k,t+1}$ is uploaded to BS by the wireless channel.

3) Global aggregation and model updating: The local FL models from all users are aggregated according to Eq. (10) in ES. Afterward, according to the variation of the SNR between the users and BS, we calculate the new pruning proportion ζ_{t+1}. When $\zeta_{t+1} < \zeta_t$, more weights of the global FL model should be pruned; otherwise, add weights randomly until $\zeta_{t+1} = \zeta_t$. Finally, the pruned global FL model $\hat{\mathbf{w}}_{g,t+1}$ is downloaded to each user to update their local FL models:

$$\hat{\mathbf{w}}_{k,t+1} = \hat{\mathbf{w}}_{g,t+1}. \tag{27}$$

Assuming the total number of communication rounds is T, the proposed PSFL assisted by AGP is summarized in **Algorithm 2**.

Algorithm 2 PSFL assisted by AGP

Input: T, $\psi_{k,t}$.
Output: $\hat{\mathbf{w}}_{g,T}$.

```
for each communication round t ∈ T do
    BS do
    Aggregate local FL models from users according to Eq. (10).
    Calculate the pruning proportion ζ_t by Eq. (26).
    Prune the global FL model w_{g,t} and obtain ŵ_{g,t}.
    Broadcast the pruned global FL model ŵ_{g,t} to each device to update the local FL model according to Eq. (27).
    Each user do
    Train the mentor and local FL models by the PLD strategy in Algorithm 1.
    Upload the latest local FL model ŵ_{k,t+1} to the BS.
end for
```

In summary, the AGP algorithm introduces a communication-aware optimization that adjusts the size of the global model according to the real-time SNR of the transmission channel. By selectively pruning model parameters before transmission, AGP significantly reduces communication overhead and energy consumption while maintaining model accuracy. This makes AGP particularly well-suited for dynamic wireless environments where link conditions fluctuate, enhancing the robustness and efficiency of federated model aggregation and broadcast.

IV. Complexity Analysis

The average data size of each user is denoted as $D = \frac{1}{K}\sum_{k\in\mathcal{K}} N_k$. We assume the complexity of communication and computation is linearly proportional to the model size [30]. Then, the complexity analysis of PSFL in terms of communication and computation is performed.

Communication complexity analysis: In the traditional FL, without PLD and AGP, the mentor model performs both local training and global aggregation. Hence, the communication complexity is $O(T|\mathbf{w}'_{k,t}|)$, where $|\cdot|$ represents the operator of calculating the size of parameter weights. In PSFL, the complexity of communication is $O(T|\hat{\mathbf{w}}_{k,t}|\zeta_t)$, which is much smaller than traditional FL for $|\hat{\mathbf{w}}_{k,t}| < |\mathbf{w}_{k,t}| \ll |\mathbf{w}'_{k,t}|$ and $\zeta_t < 1$.

Computation complexity analysis: The computation complexity that directly learning the large mentor model in FL is $O(TD|\mathbf{w}'_{k,t}|)$. In PSFL, the computation complexity consists of two parts, i.e., local mentor and FL models training, which are $O(TD|\mathbf{w}'_{k,t}|)$ and $O(TD|\hat{\mathbf{w}}_{k,t}|)$, respectively. Hence, the computation complexity of PSFL is $O(TD|\mathbf{w}'_{k,t}|) + O(TD|\hat{\mathbf{w}}_{k,t}|)$. Since the model size of the pruned FL model is much smaller than the mentor model, the extra computation cost of learning the FL model is much smaller than learning the large mentor model, namely, the computation complexity of PSFL is $O(TD|\mathbf{w}'_{k,t}|)$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCCN.2025.3586904, IEEE Transactions on Cognitive Communications and Networking

In practice, compared with the standard FedAvg algorithm, the extra computation cost of training the FL model is much smaller than training the large mentor model with the PLD strategy. AGP can also stably reduce communication energy consumption due to the adaptive pruning. Thus, the proposed PSFL is much more communication-efficient than the standard FedAvg algorithm, and meanwhile does not introduce much computation cost.

V. Numerical Results

A. Simulation Settings

Firstly, we assess the proposed PSFL scheme using the MNIST [31], Fashion-MNIST [32], CIFAR-10 [33], and CIFAR-100 [34] datasets. In the MNIST and Fashion-MNIST datasets, the training set contains 60,000 samples and the testing set contains 10,000 samples, distributed across 10 categories. The CIFAR-10 dataset comprises 50,000 RGB images as training samples and 10,000 RGB images as test samples, distributed across 10 categories. The CIFAR-100 dataset comprises 50,000 RGB images as training samples and 10,000 RGB images as test samples, distributed across 100 categories. Furthermore, we apply a Dirichlet distribution [35] to generate the non-IID data partition among users, where the concentration parameter of the Dirichlet distribution is denoted as r and set to 0.9 by default. Thus, each device can have relatively balanced data samples of some classes.

Secondly, we assume the participation of 9 users as clients in the training process. The maximum transmit power is configured as $P_{k,\max} = 0.1$ W. The maximum and minimum SNR between the users and BS are set as $\psi_{\max} = 25$ dB and $\psi_{\min} = 0$ dB, respectively. The AWGN channel is used as the simulated physical channel for the GSC model. The SNRs between the BS and clients fluctuate across different communication rounds, ranging from 0 dB to 25 dB.

Thirdly, Masked Autoencoders (MAE) [36] is a well-known large generative model, which utilizes ViT as the encoder-decoder to learn accurate semantic representation of images and perform high-quality construction of images. Therefore, we adopt the MAE to construct the semantic encoder and decoder in the GSC model, respectively. As shown in TABLE I, three kinds of GSC models, denoted as GSC-M, GSC-L, and GSC-H, are used as the mentor models. A ResNet-18 [25] is used as the semantic encoder, and three layers of a deconvolutional network construct the semantic decoder. As a result, each client is equipped with a unified student model and a personalized mentor model based on the size of their local datasets.

Finally, our simulations are performed with the PyTorch framework on a server, which has an Intel Xeon CPU (2.4 GHz, 128 GB RAM) and an NVIDIA A800 GPU (80 GB SGRAM).

TABLE I: Data distribution and local model settings

Client	C0	C1	C2	C3	C4	C5	C6	C7	C8
Local data vol.	1854	3703	4429	4783	5467	5634	5891	8958	9281
Mentor model	GSC-M				GSC-L			GSC-H	
Parameters (mentor)	112 M				330 M			658 M	
Student model	CSC								
Parameters (student)	54.7 M								

B. Evaluation of the proposed GSC model

To emphasize the advantages of the GAI architecture intuitively, we display some restructured images generated by the GSC and CSC models. Additionally, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are the metrics to quantify the quality of the reconstructed images. PSNR measures the quality of a reconstructed image, typically expressed in decibels, with higher values indicating better image quality. The definition of PSNR is as follows:

$$\mathrm{PSNR}(\mathbf{m}, \hat{\mathbf{m}}) = 10 \cdot \log_{10} \left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}(\mathbf{m}, \hat{\mathbf{m}})} \right), \tag{28}$$

where MAX_I denotes the maximum possible pixel value of the image, which is typically 255 for an 8-bit image. Similarly, SSIM is a metric that gauges the perceived similarity between two images, factoring in three key components - luminance, contrast, and structure. The definition of SSIM is outlined as follows:

$$\mathrm{SSIM}(\mathbf{m}, \hat{\mathbf{m}}) = \frac{(2\varphi_{\mathbf{m}}\varphi_{\hat{\mathbf{m}}} + c_1)(2\phi_{\mathbf{m}\hat{\mathbf{m}}} + c_2)}{(\varphi_{\mathbf{m}}^2 + \varphi_{\hat{\mathbf{m}}}^2 + c_1)(\phi_{\mathbf{m}}^2 + \phi_{\hat{\mathbf{m}}}^2 + c_2)}, \tag{29}$$

where $\varphi_{\mathbf{m}}$ and $\varphi_{\hat{\mathbf{m}}}$ are their means; $\phi_{\mathbf{m}}^2$ and $\phi_{\hat{\mathbf{m}}^2}$ are their variances; $\phi_{\mathbf{m}\hat{\mathbf{m}}}$ is their covariance; c_1 and c_2 are two constants used to avoid division by zero.

Fig. 4 presents the evaluation results on the MNIST dataset, where both Fig. 4(b) and Fig. 4(c) reconstruct the content of Fig. 4(a). However, Fig. 4(b) provides superior image details. The values shown above Fig. 4(b) and 4(c) indicate the differences in PSNR and SSIM when compared to Fig. 4(a), with Fig. 4(b) attaining higher scores. Fig. 5 shows the evaluation results on the Fashion-MNIST dataset, where both Fig. 5(b) and Fig. 5(c) reconstruct the content of Fig. 5(a). However, Fig. 5(b) is noticeably clearer than Fig. 5(c). The PSNR and SSIM results further confirm the higher quality of Fig. 5(b). Fig. 6 and 7 illustrate the evaluation results on the CIFAR-10 and CIFAR-100 datasets. In both cases, Fig. 6(b) and 7(b) accurately reconstruct the original images, while the reconstructions in Fig. 6(c) and 7(c) appear blurry. The PSNR and SSIM results also indicate that the images generated by the GSC models are of higher quality than those produced by the CSC models.

The superior performance of the GSC model is attributed to the advantages of the GAI model architecture (i.e., MAE), which extracts more precise semantic information compared to CNN architectures. Furthermore, due to its robust generative capability, the proposed GSC model achieves more accurate image reconstruction in SC.

C. Evaluation of the proposed PSFL

This subsection evaluates the proposed PSFL scheme's performance in terms of loss and accuracy. Note that accuracy

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCCN.2025.3586904, IEEE Transactions on Cognitive Communications and Networking

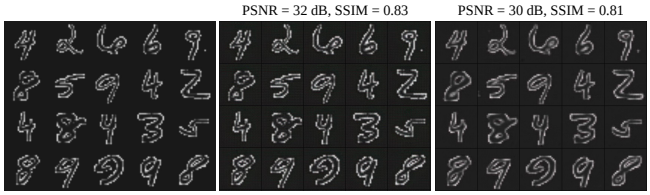


Fig. 4: Image transmission results on the MNIST dataset. (a) Original images. (b) Reconstructed images based on the GSC model. (c) Reconstructed images based on the CSC model.

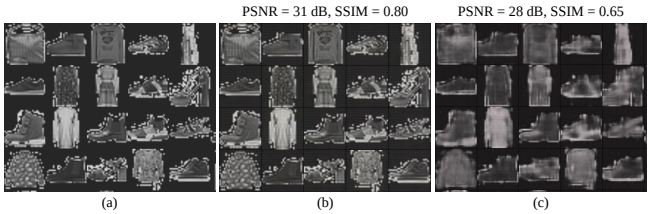


Fig. 5: Image transmission results on the Fashion-MNIST dataset. (a) Original images. (b) Reconstructed images based on the GSC model. (c) Reconstructed images based on the CSC model.

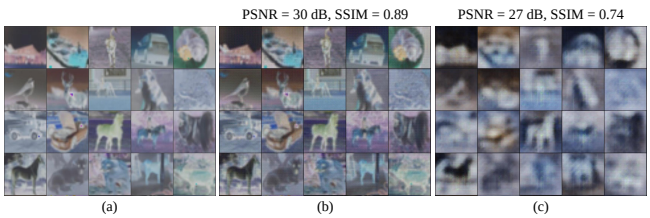


Fig. 6: Image transmission results on the CIFAR-10 dataset. (a) Original images. (b) Reconstructed images based on the GSC model. (c) Reconstructed images based on the CSC model.

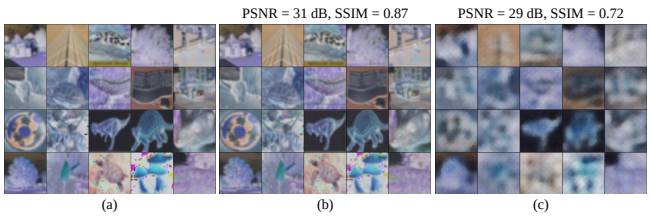


Fig. 7: Image transmission results on the CIFAR-100 dataset. (a) Original images. (b) Reconstructed images based on the GSC model. (c) Reconstructed images based on the CSC model.

refers to the probability of correctly classifying an image reconstructed by the GSC model using a pre-trained classifier network, which can measure the scheme's performance from a semantic perspective. We employ ResNet-101 [37] as the classifier and train it on four datasets, thus obtaining the corresponding pre-trained weights. The training result is shown in Fig. 8, which shows that the ResNet-101 has achieved good accuracy. This ensures that the pretrained ResNet-101 can accurately distinguish between generated and real images, thereby effectively guiding the training of the GSC model.

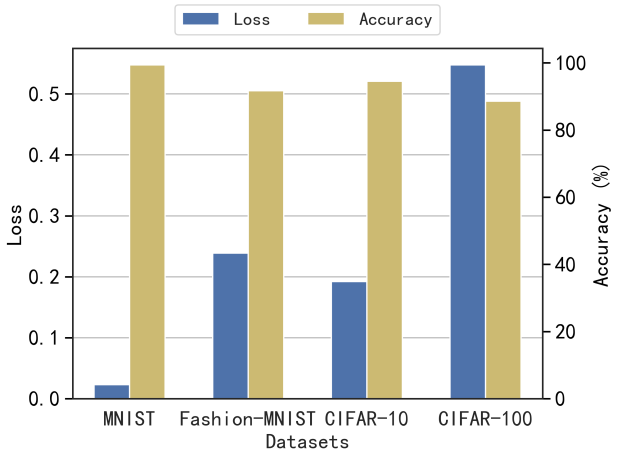


Fig. 8: Training results of the ResNet-101 on the four datasets.

Fig. 9 and 10 present the training results of the mentor and student models across four datasets. Note that the loss and accuracy results shown in the figures represent the mean loss and accuracy of the GSC model across all clients. We can see that the mentor model consistently achieves lower loss and higher accuracy compared to the student model, highlighting its strong capability to guide the student model's learning process. Specifically, for the MNIST and Fashion-MNIST datasets, as shown in Fig. 9(a)-(b) and Fig. 10(a)-(b), the simplicity of these datasets enables the student model to achieve performance levels close to those of the mentor model. This indicates that the student model effectively learns from the mentor during the training process. In contrast, for the more complex CIFAR-10 and CIFAR-100 datasets, as depicted in Fig. 9(c)-(d) and Fig. 10(c)-(d), the mentor model demonstrates a significant performance advantage, with faster loss convergence and consistently higher accuracy. This underscores the mentor model's ability to provide effective guidance in handling more challenging tasks. Overall, these results demonstrate that within the PSFL framework, both the mentor and student models can continually learn and improve, ensuring the effectiveness of information exchange during the training process.

Secondly, we evaluate the model performance under different r. Fig. 11 and Fig. 12 illustrate the training results under different Dirichlet distributions on four datasets. The results suggest that the performance of the model becomes worse with the decrease in r, as a smaller r results in greater differences

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCCN.2025.3586904, IEEE Transactions on Cognitive Communications and Networking

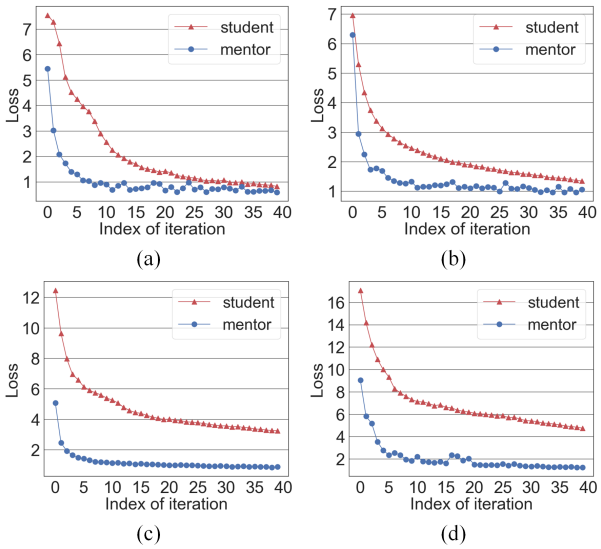


Fig. 9: Loss versus iteration under student and mentor models on datasets (a) MNIST, (b) Fashion-MNIST, (c) CIFAR-10, and (d) CIFAR-100.

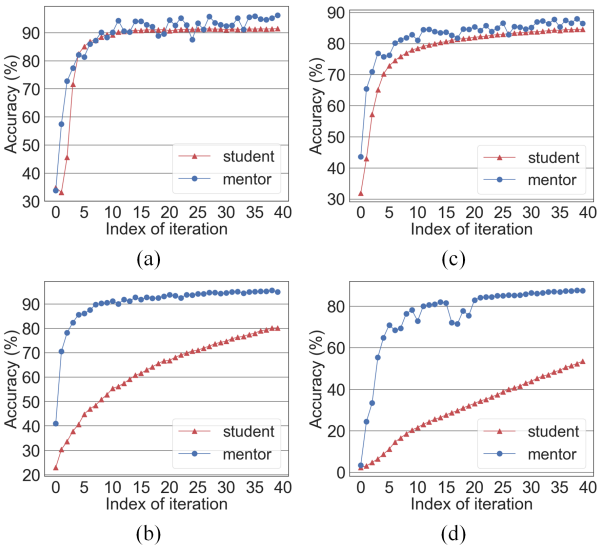


Fig. 10: Accuracy versus iteration under student and mentor models on datasets (a) MNIST, (b) Fashion-MNIST, (c) CIFAR-10, and (d) CIFAR-100.

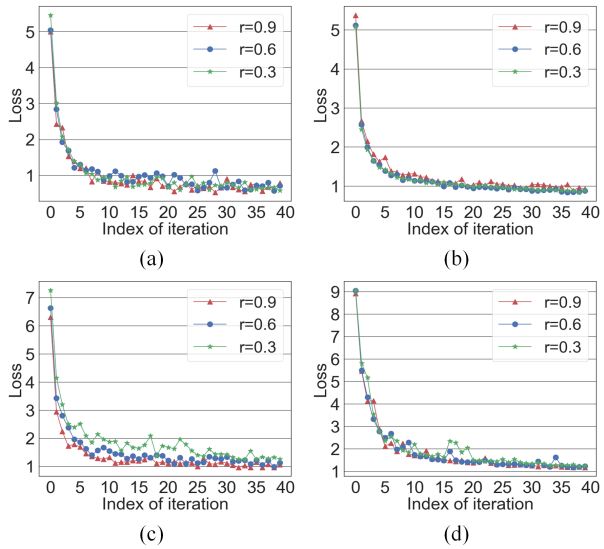


Fig. 11: Loss versus iteration under different concentration parameters r of the Dirichlet distribution on datasets (a) MNIST, (b) Fashion-MNIST, (c) CIFAR-10, and (d) CIFAR-100.

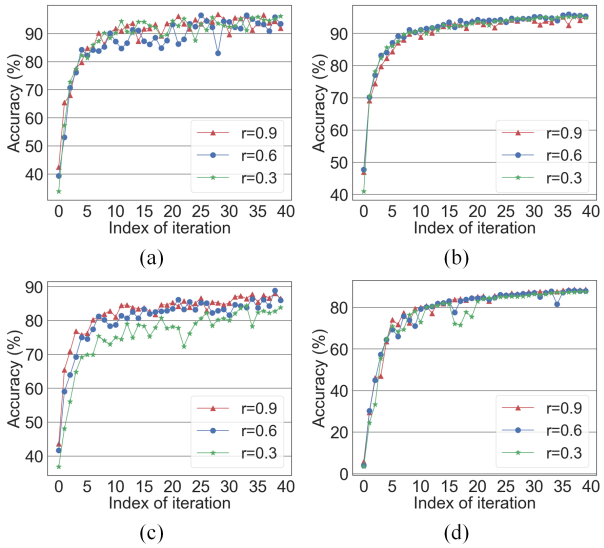


Fig. 12: Accuracy versus iteration under different concentration parameters r of the Dirichlet distribution on datasets (a) MNIST, (b) Fashion-MNIST, (c) CIFAR-10, and (d) CIFAR-100.

between different client data. Specifically, as shown in Fig. 11(a)-(b) and Fig. 12(a)-(b), for the simpler MNIST and Fashion-MNIST datasets, the differences are smaller under varying r. For the more complex CIFAR-10 and CIFAR-100 datasets, as shown in Fig. 11(c)-(d) and Fig. 12(c)-(d), the increase in r has a more significant impact on the FL model. This means that the proposed PSFL can not perform well with non-IID data when the dataset is complex, which could lead to further improvements in the FL scheme in the future.

Lastly, ablation experiments are performed to evaluate the functions of PLD and AGP in the PSFL scheme. The results are shown in Fig. 13 and Fig. 14. Note in the PSFL without PLD, each user selects the GSC-M model as the mentor. Fig. 13(a)-(b) and Fig. 14(a)-(b) show that the PSFL without AGP, PSFL, and PSFL without PLD schemes have similar performance. This may be due to the simplicity of the MNIST and Fashion-MNIST datasets. In Fig. 13(c)-(d) and Fig. 14(c)-(d), the PSFL without AGP achieves the lowest loss and the highest accuracy, while the PSFL without PLD has the worst result. We speculate that AGP reduces the parameters of the model but impacts the performance, whereas PLD effectively improves model accuracy.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCCN.2025.3586904, IEEE Transactions on Cognitive Communications and Networking

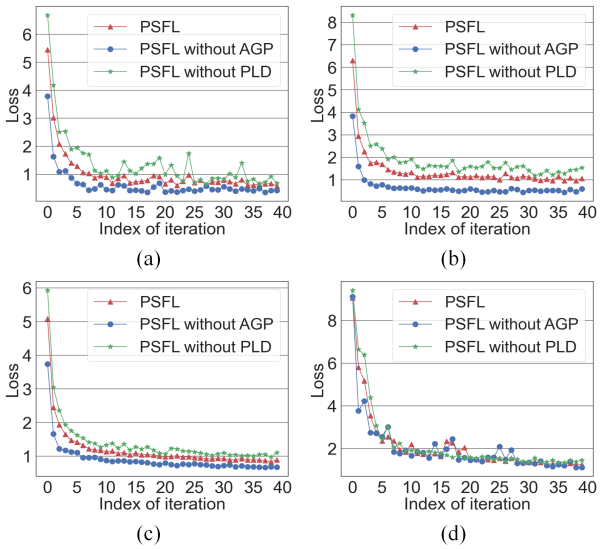


Fig. 13: Loss versus iteration under different methods on datasets (a) MNIST, (b) Fashion-MNIST, (c) CIFAR-10, and (d) CIFAR-100.

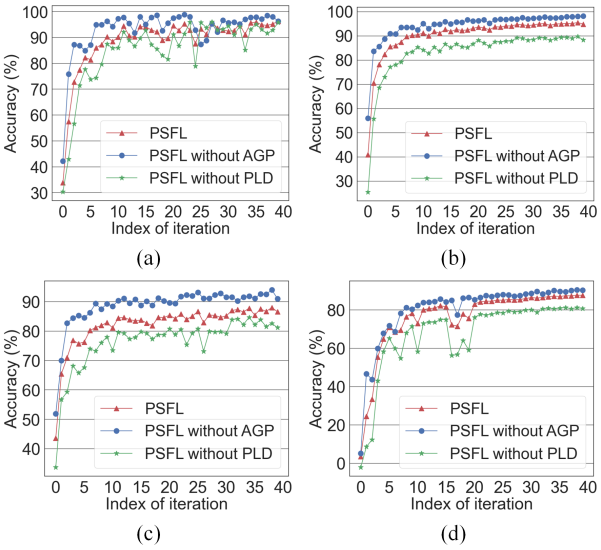


Fig. 14: Accuracy versus iteration under different methods on datasets (a) MNIST, (b) Fashion-MNIST, (c) CIFAR-10, and (d) CIFAR-100.

D. Evaluation of different contenders

This subsection compares the proposed PSFL to the other FL schemes in terms of global loss, global accuracy, and local accuracy. The following methods are introduced in this experiment as contenders:

- FedAvg: A common FL approach, which is equivalent to the PSFL without PLD and AGP algorithms [15].
- STC: A compressed FL framework that is designed to meet the requirements of the FL environment [38].
- FTTQ: A parameter quantization-based communication-efficient FL approach [39].
- FedPAQ: A communication-efficient FL method with periodic averaging and quantization [40].
- PSFL: The proposed FL approach in this paper.

Except for PSFL, the other schemes adopt the GSC-M as the FL model. Additionally, for simplicity, we evaluate these methods on the Fashion-MNIST and CIFAR-10 datasets and set $r = 0.9$. Fig. 15 and Fig. 16 show evaluation results.

The loss results in Fig. 15(a) and Fig. 16(a) suggest that the proposed PSFL could converge to the best point on the Fashion-MNIST and CIFAR-10 datasets, while FedPAQ performs the worst. Additionally, the FedAvg and FTTQ schemes perform better, while STC performs worse. Fig. 15(b) and Fig. 16(b) suggest that the accuracy of the global FL model obtained by our method is the best and is significantly better than that of the other contenders. The FedAvg method performs better than the FTTQ, STC, and FedPAQ methods, with FedPAQ performing the worst. Fig. 15(c) and Fig. 16(c) display that the proposed PSFL enables all local FL models to achieve the best final accuracy. The performance of FedAvg and FTTQ is only slightly worse than ours, while STC performs only slightly better than FedPAQ, with FedPAQ results being the worst on both datasets.

We speculate that the excellent accuracy performance of the PSFL is mainly attributed to the PLD strategy. PLD allows clients to freely select the most compatible GSC models, thus fully utilizing available resources and achieving high accuracy. Hence, with the proposed PSFL, all clients could achieve the best performance despite differences in their local data resources. Moreover, without transmitting all parameter weights, the AGP algorithm also ensures effective model information exchange among clients, thereby maintaining the accuracy of the GSC models.

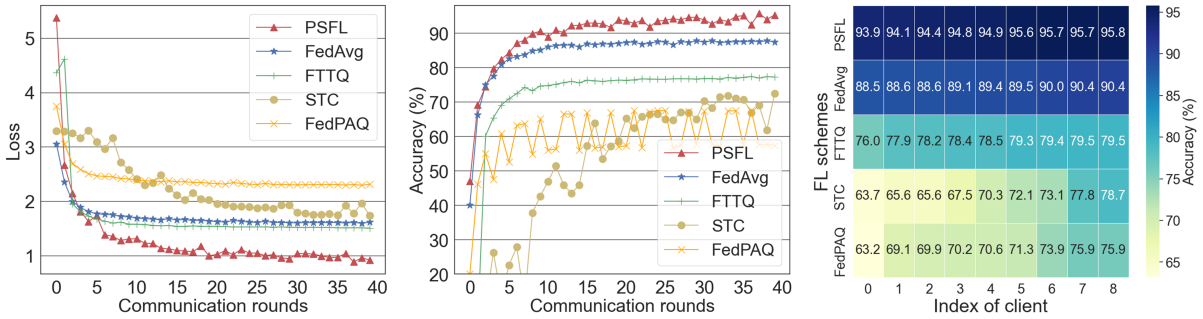
E. Performance Evaluation of Communication Energy Consumption

This subsection evaluates the performance of the proposed PSFL and other schemes in terms of communication energy consumption. Fig. 17 shows the communication energy consumption of each communication round using different FL schemes.

Fig. 17 shows that the boxplot of the proposed PSFL is at the bottom, which means the communication energy consumption of PSFL is the lowest. Meanwhile, the boxplot of PSFL is the flattest, namely, the energy consumption change of each round is the smallest under dynamic SNR. Similarly, FedAvg has the highest energy consumption and the most dramatic variation in energy consumption.

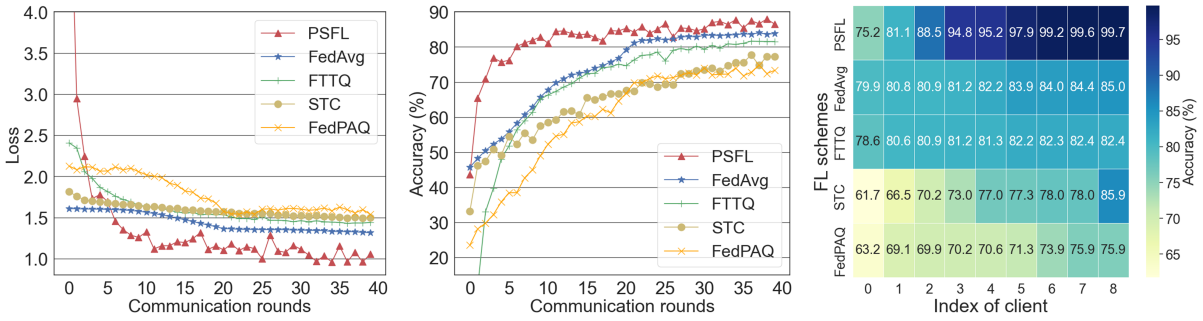
Hence, we demonstrate that the proposed PSFL can ensure low communication energy consumption in dynamic SNR. The low and stable communication energy consumption of the PSFL can be attributed to AGP. The AGP algorithm prunes the FL model while considering the dynamic SNR, thus reducing the communication energy consumption efficiently in wireless communications. Furthermore, the pruning operator is processed only on the server, which could incur no extra cost to clients.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCCN.2025.3586904, IEEE Transactions on Cognitive Communications and Networking



(a) (b) (c)

Fig. 15: Comparison results of different schemes on Fashion-MNIST in terms of (a) global loss, (b) global accuracy, and (c) local accuracy of each client.



(a) (b) (c)

Fig. 16: Comparison results of different schemes on CIFAR-10 in terms of (a) global loss, (b) global accuracy, and (c) local accuracy of each client.

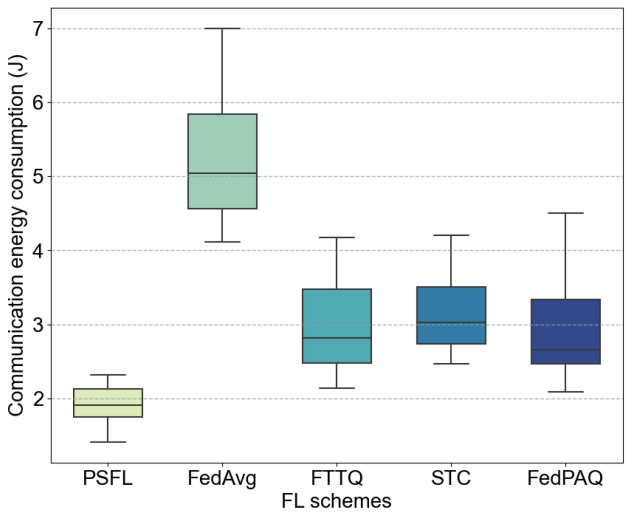


Fig. 17: Communication energy consumption of each communication round using different methods.

VI. Conclusion

In this paper, we propose a novel GSC model that leverages the strengths of GAI to enhance the performance of SC between users and the BS. Then, the PSFL framework is presented to enable users and the BS to collaboratively train GSC models while accommodating the training requirements of heterogeneous users. This framework first introduces the PLD strategy during the local training phase, in which each user selects a suitable GSC model as a mentor and a unified CSC model as a student. The two models engage in mutual learning based on KD. After local training, the unified CSC model is utilized as the local FL model and uploaded to the BS for parameter aggregation, thereby obtaining the global FL model. Secondly, PSFL applies the AGP algorithm in the global aggregation phase, which prunes the aggregated global FL model according to the real-time SNR. The AGP algorithm reduces the transmitted model parameters and achieves the trade-off between the communication energy and model accuracy. Finally, numerical results demonstrate the feasibility and efficiency of the proposed PSFL.

In the future, we will work on improving the performance of the proposed PSFL on non-IID data by introducing the latest personalized FL algorithms. Additionally, since the parameters of the CSC model may relate to user privacy, improving the security of model parameters during parameter aggregation is also a potential issue.

References

[1] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, and X. You, "Large AI model-based semantic communications," *IEEE Wireless Communications*, vol. 31, no. 3, pp. 68–75, 2024.

[2] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 93–99, 2019.
[3] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.
[4] J. Wang, S. Wang, J. Dai, Z. Si, D. Zhou, and K. Niu, "Perceptual learned source-channel coding for high-fidelity image semantic transmission," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 2022, pp. 3959–3964.
[5] T. Han, Q. Yang, Z. Shi, S. He, and Z. Zhang, "Semantic-preserved communication system for highly efficient speech transmission," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 245–259, 2023.
[6] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, D. Niyato, and O. A. Dobre, "Large language model enhanced multi-agent systems for 6g communications," *IEEE Wireless Communications*, pp. 1–8, 2024.
[7] L. Dong, F. Jiang, Y. Peng, K. Wang, K. Yang, C. Pan, and R. Schober, "Lambo: Large AI model empowered edge intelligence," *IEEE Communications Magazine*, vol. 63, no. 4, pp. 88–94, 2025.
[8] F. Jiang, C. Tang, L. Dong, K. Wang, K. Yang, and C. Pan, "Visual language model-based cross-modal semantic communication systems," *IEEE Transactions on Wireless Communications*, vol. 24, no. 5, pp. 3937–3948, 2025.
[9] F. Jiang, L. Dong, Y. Peng, K. Wang, K. Yang, C. Pan, and X. You, "Large AI model empowered multimodal semantic communications," *IEEE Communications Magazine*, vol. 63, no. 1, pp. 76–82, 2025.
[10] H. Du, J. Wang, D. Niyato, J. Kang, Z. Xiong, and D. I. Kim, "Ai-generated incentive mechanism and full-duplex semantic communications for information sharing," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 9, pp. 2981–2997, 2023.
[11] Y. Lin, H. Du, D. Niyato, J. Nie, J. Zhang, Y. Cheng, and Z. Yang, "Blockchain-aided secure semantic communication for ai-generated content in metaverse," *IEEE Open Journal of the Computer Society*, vol. 4, pp. 72–83, 2023.
[12] S. Guo, Y. Wang, S. Li, and N. Saeed, "Semantic importance-aware communications using pre-trained language models," *IEEE Communications Letters*, vol. 27, no. 9, pp. 2328–2332, 2023.
[13] L. Dong, Y. Peng, F. Jiang, K. Wang, and K. Yang, "Explainable semantic federated learning enabled industrial edge network for fire surveillance," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 12, pp. 14 053–14 061, 2024.
[14] B. Rao, J. Zhang, D. Wu, C. Zhu, X. Sun, and B. Chen, "Privacy inference attack and defense in centralized and federated learning: A comprehensive survey," *IEEE Transactions on Artificial Intelligence*, 2024.
[15] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 13, no. 3, pp. 1–207, 2019.
[16] M.-D. Nguyen, S.-M. Lee, Q.-V. Pham, D. T. Hoang, D. N. Nguyen, and W.-J. Hwang, "Hcfl: A high compression approach for communication-efficient federated learning in very large scale iot networks," *IEEE Transactions on Mobile Computing*, pp. 1–13, 2022.
[17] Y. Wang, L. Lin, and J. Chen, "Communication-efficient adaptive federated learning," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 22 802–22 838. [Online]. Available: https://proceedings.mlr.press/v162/wang22o.html
[18] R. Hönig, Y. Zhao, and R. Mullins, "DAdaQuant: Doubly-adaptive quantization for communication-efficient federated learning," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 8852–8866. [Online]. Available: https://proceedings.mlr.press/v162/honig22a.html
[19] L. Dong, F. Jiang, and Y. Peng, "Attention-based uav trajectory optimization for wireless power transfer-assisted iot systems," *IEEE Transactions on Industrial Electronics*, pp. 1–9, 2025.
[20] L. Dong, F. Jiang, M. Wang, Y. Peng, and X. Li, "Deep progressive reinforcement learning-based flexible resource scheduling framework for irs and uav-assisted mec system," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 2, pp. 2314–2326, 2025.
[21] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," *Proceedings of the National Academy of Sciences*, vol. 118, no. 17, p. e2024789118, 2021.
[22] F. Jiang, C. Pan, L. Dong, K. Wang, M. Debbah, D. Niyato, and Z. Han, "A comprehensive survey of large ai models for future communications: Foundations, applications and challenges," *arXiv preprint arXiv:2505.03556*, 2025.
[23] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, and X. You, "Large generative model assisted 3d semantic communication," *arXiv preprint arXiv:2403.05783*, 2024.
[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[25] E. Rezende, G. Ruppert, T. Carvalho, F. Ramos, and P. De Geus, "Malicious software classification using transfer learning of resnet-50 deep neural network," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2017, pp. 1011–1014.
[26] Y. Qu, M. P. Uddin, C. Gan, Y. Xiang, L. Gao, and J. Yearwood, "Blockchain-enabled federated learning: A survey," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–35, 2022.
[27] T. Shen, J. Zhang, X. Jia, F. Zhang, G. Huang, P. Zhou, K. Kuang, F. Wu, and C. Wu, "Federated mutual learning," *arXiv preprint arXiv:2006.16765*, 2020.
[28] A. Jafari, M. Rezagholizadeh, P. Sharma, and A. Ghodsi, "Annealing knowledge distillation," *arXiv preprint arXiv:2104.07163*, 2021.
[29] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. Ieee, 2018, pp. 1–2.
[30] C. Wu, F. Wu, L. Lyu, Y. Huang, and X. Xie, "Communication-efficient federated learning via knowledge distillation," *Nature communications*, vol. 13, no. 1, pp. 1–8, 2022.
[31] 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.
[32] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.
[33] P. Zhang, C. Wang, C. Jiang, and Z. Han, "Deep reinforcement learning assisted federated learning algorithm for data management of iiot," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 12, pp. 8475–8484, 2021.
[34] J. Huang, L. Ye, and L. Kang, "Fedsr: A semi-decentralized federated learning algorithm for non-iidness in iot system," *arXiv preprint arXiv:2403.14718*, 2024.
[35] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 713–10 722.
[36] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.
[37] Y. Jusman, "Comparison of prostate cell image classification using cnn: Resnet-101 and vgg-19," in *2023 IEEE 13th International Conference on Control System, Computing and Engineering (ICCSCE)*. IEEE, 2023, pp. 74–78.
[38] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 9, pp. 3400–3413, 2019.
[39] J. Xu, W. Du, Y. Jin, W. He, and R. Cheng, "Ternary compression for communication-efficient federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
[40] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2021–2031.

BIOGRAPHIES

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCCN.2025.3586904, IEEE Transactions on Cognitive Communications and Networking

Yubo Peng received his B.S. and M.S. degrees from Hunan Normal University, Changsha, China, in 2019 and 2024. He is pursuing a doctor's degree from the School of Intelligent Software and Engineering at Nanjing University. His main research interests include semantic communication and large models.

Feibo Jiang received his B.S. and M.S. degrees in the School of Physics and Electronics from Hunan Normal University, China, in 2004 and 2007, respectively. He received his Ph.D. degree in the School of Geosciences and Info-physics from the Central South University, China, in 2014. He is currently an associate professor at the Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, China. His research interests include artificial intelligence, fuzzy computation, Internet of Things, and mobile edge computing.

Li Dong received the B.S. and M.S. degrees in the School of Physics and Electronics from Hunan Normal University, China, in 2004 and 2007, respectively. She received her Ph.D. degree in the School of Geosciences and Info-physics from the Central South University, China, in 2018. Her research interests include machine learning, Internet of Things, and mobile edge computing.

Kezhi Wang received the Ph.D. degree in Engineering from the University of Warwick, U.K. He was with the University of Essex and Northumbria University, U.K. Currently, he is a Professor with the Department of Computer Science, Brunel University London, U.K. His research interests include wireless communications, mobile edge computing, and machine learning.

Kun Yang received his PhD from the Department of Electronic & Electrical Engineering of University College London (UCL), UK. He is currently a Chair Professor of Nanjing University and also an affiliated professor at the University of Essex. His main research interests include wireless networks and communications, communication-computing cooperation, and new AI (artificial intelligence) for wireless. He has published 500+ papers and filed 50 patents. He serves on the editorial boards of a number of IEEE journals (e.g., IEEE WCM, TVT, TNB). He is a Deputy Editor-in-Chief of IET Smart Cities Journal. He has been a Judge of the GSMA GLOMO Award at World Mobile Congress – Barcelona since 2019. He was a Distinguished Lecturer of IEEE ComSoc (2020-2021), a Recipient of the 2024 IET Achievement Medals, and a Recipient of the 2024 IEEE CommSoft TC's Technical Achievement Award. He is a Member of Academia Europaea (MAE), a Fellow of IEEE, a Fellow of IET, and a Distinguished Member of ACM.