SANSee: A Physical-layer Semantic-aware Networking Framework for Distributed Wireless Sensing

Huixiang Zhu, Yong Xiao, *Senior Member, IEEE*, Yingyu Li, Guangming Shi, *Fellow, IEEE*, and Marwan Krunz, *Fellow, IEEE*

◆

Abstract—Contactless device-free wireless sensing has recently attracted significant interest due to its potential to support a wide range of immersive human-machine interactive applications using ubiquitously available radio frequency (RF) signals. Traditional approaches focus on developing a single global model based on a combined dataset collected from different locations. However, wireless signals are known to be location and environment specific. Thus, a global model results in inconsistent and unreliable sensing results. It is also unrealistic to construct individual models for all the possible locations and environmental scenarios. Motivated by the observation that signals recorded at different locations are closely related to a set of physical-layer semantic features, in this paper we propose SANSee, a semantic-aware networking-based framework for distributed wireless sensing. SANSee allows models constructed in one or a limited number of locations to be transferred to new locations without requiring any locally labeled data or model training. SANSee is built on the concept of physical-layer semantic-aware network (pSAN), which characterizes the semantic similarity and the correlations of sensed data across different locations. A pSAN-based zero-shot transfer learning solution is introduced to allow receivers in new locations to obtain location-specific models by directly aggregating the models trained by other receivers. We theoretically prove that models obtained by SANSee can approach the locally optimal models. Experimental results based on real-world datasets are used to verify that the accuracy of the transferred models obtained by SANSee matches that of the models trained by the locally labeled data based on supervised learning approaches.

1 INTRODUCTION

Wireless sensing has recently attracted significant interest due to its potential to achieve device-free movement detection and tracking in a wide range of applications, including smart healthcare, urban sensing, and unmanned surveillance systems. It is a key enabler of emerging applications that require immersive contact-free human-machine interactions, including augmented reality/virtual reality (AR/VR) and Tactile Internet [2], [3]. Recent results show that by detecting changes in the RF signal propagation and reflection patterns caused by the human body, it is possible to recognize a wide range of human actions and gestures, such as falling, walking, sitting, etc. Furthermore, if wireless sensing data collected by multiple receivers can be jointly analyzed, more fine-grained human gestures, such as hand gestures and finger movement, can be detected [4].

H. Zhu is with the School of Electronic Information and Communications at the Huazhong University of Science and Technology, Wuhan, China 430074 (e-mail: zhuhuixiang@hust.edu.cn).

Y. Xiao is affiliated with the School of Electronic Information and Communications at the Huazhong University of Science and Technology, Wuhan 430074, China, the Peng Cheng Laboratory, Shenzhen, Guangdong 518055, China, and the Pazhou Laboratory (Huangpu), Guangzhou, Guangdong 510555, China (e-mail: yongxiao@hust.edu.cn).

Y. Li is with the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan, China 430074 (e-mail: liyingyu29@cug.edu.cn).

G. Shi is with the Peng Cheng Laboratory, Shenzhen, Guangdong 518055, China, the School of Artificial Intelligence, the Xidian University, Xi'an, Shaanxi 710071, China, and the Pazhou Laboratory (Huangpu), Guangzhou, Guangdong 510555, China (e-mail: gmshi@xidian.edu.cn).

M. Krunz is with the Department of Electrical and Computer Engineering, the University of Arizona, Tucson, AZ 85721 (e-mail: krunz@arizona.edu).

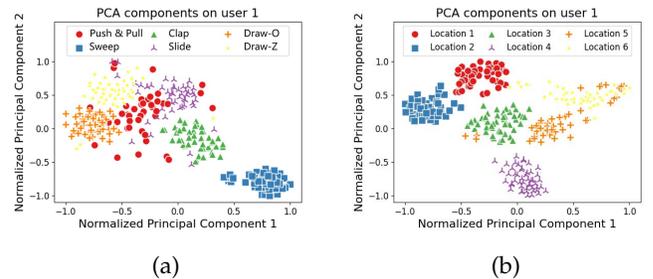Fig. 1: (a) Visualization of the statistical features of wireless signals recorded at the same location when different human gestures are performed, and (b) visualization of statistical diversity of wireless signals recorded by receivers at different locations when the same gesture is performed.

Most existing works on wireless sensing adopt a one-fits-all approach, in which a centralized model is

trained based on wireless sensing data recorded from a few locations and applied to a much wider range of locations and environments. However, wireless signals are known to exhibit highly temporal and spatial heterogeneity. Specifically, wireless signals are highly dependent on location, environment, and human-related factors. For example, different locations of transmitters, receivers, and objects, as well as room layouts will result in drastically different signal characteristics and data distributions. Furthermore, different body movement patterns (e.g., human gestures) and orientations will also result in different spatial and temporal variations of wireless sensing data. To shed more light on this observation, in Fig. 1 we use principal components analysis (PCA) to reduce data dimension and then visualize the resulting 2-dimension statistical features of wireless sensing signals [5], i.e., channel state information (CSI), recorded when the same person performs different gestures at the same location (Fig. 1(a)) and when the person performs the same gesture at different locations (Fig. 1(b)). We can observe that the statistical features vary significantly when different gestures are performed or when receivers are deployed at different locations. Accordingly, training a single global model by combining sensing data collected at different locations, while ignoring the unique features of each individual location, environment, and gesture profile, will significantly reduce the wireless sensing accuracy and will result in highly unreliable sensing performance across different locations and gestures.

One possible solution is to train separate models for different locations and environments. Unfortunately, this approach incurs too much overload and relies on a large number of high-quality labeled data samples. Also, due to physical space and cost limitations, it is generally unrealistic to have a highly dense deployment of sensors and receivers to collect data that covers all spatial and temporal resolutions of different users and their gestures. To summarize, due to the heterogeneity of wireless signals and the scarcity of lablled samples, it is difficult for conventional distributed wireless sensing solutions to achieve a desired gesture recognition accuracy, especially when most receivers cannot collect labeled data samples or construct local models due to their limited computational and storage capabilities.

To overcome the above challenges, we propose SANSee, a distributed wireless sensing framework that transfers the gesture recognition models trained for one or a few locations to new locations without training new models or collecting new data samples. Our proposed model is motivated by the observation that the statistics of the wireless signals recorded in a given location are closely related to a set of physical-layer semantic features, such as the spatial layout, environmental features, and gesture profiles. These physical-layer semantic features can be utilized to infer the statistical correlations between wireless sensing signals across different locations and environments for location-specific model construction and transfer. More specifically, we develop a novel *physical-layer semantic-aware networking* (pSAN) framework to characterize the similarity between physical-layer semantic features and correlations of wireless signal distributions at different locations and

environmental scenarios. We then propose a *pSAN-based zero-shot transfer learning solution*, in which receivers at new locations and environments obtain location-specific gesture recognition models by directly aggregating the already trained models of other receivers. In our solution, the aggregation coefficients of the model transfer are calculated based on the correlations between semantic features of different locations. We theoretically prove that the aggregated model obtained by SANSee approaches the locally optimal model without requiring any locally labeled data or local model training. Extensive experiments conducted based on real-world datasets are presented to corroborate our theoretical results.

The key contributions of this paper are as follows:

- We identify the physical-layer semantic features, including environment-related and gesture-related semantics, called E- and G-semantics, respectively, that determine the distributions of wireless sensing signals under different physical environments and gesture profiles. We then introduce the pSAN framework, which captures similarity between physical-layer semantics of different locations at different physical environments.
- We develop a zero-shot transfer learning solution based on pSAN, which allows receivers in new locations to obtain location-specific models by linearly aggregating the models trained by a few receivers.
- We present theoretical bounds on model training error and transfer errors of SANSee. We prove that the localized models obtained by SANSee approaches the locally optimal model in each specific location even without locally labeled data or local model training.
- Extensive experiments are conducted based on real-world wireless sensing datasets consisting of multiple types of human gestures recorded at 18 different locations. Our results show that our proposed model aggregation solutions can match models trained by real labeled data, obtained through supervised learning.

The remainder of this paper is organized as follows. Related works are reviewed in Section 2. We introduce the system model and problem formulation in Section 3. An overview of SANSee framework is provided in Section 4. The detailed procedures of physical-layer semantics estimation are discussed in Section 5. The concept of semantic similarity and pSAN-based model correlation network are introduced in Section 6.1. Model training and transfer algorithms are proposed in Sections 6.2 and 6.3, respectively. Theoretical results about model training error and transfer error are derived in Section 7. Experimental results are presented in Section 8, and we conclude the paper in Section 9.

2 RELATED WORK

RF-based Wireless Sensing: Distributed wireless sensing has emerged as a promising area of research, leveraging ubiquitous wireless signals to enable contactless and

device-free localization, tracking, and activity recognition [6], [7]. Most existing works focus on capturing the spatial and temporal dynamics of a few parameters, such as Doppler frequency shift (DFS), Time-of-Flight (ToF), and Angle-of-Arrival (AoA) [8], [9]. In [10] the authors proposed SpotFi for decimeter-level human localization based on the AoA and relative ToF information of dominant incident signals from the target to multiple receivers. In [11] the authors designed a human trajectory tracking system named IndoTrack to achieve successive tracking in an indoor environment. The main idea behind IndoTrack is to first extract accurate DFS from noisy channel state information samples and then jointly estimate target velocity and location via probabilistic co-modeling of DFS and AoA information from wireless receivers. In [12] the authors proposed Widar3.0 to achieve cross-domain gesture recognition by feeding the domain-independent Body Coordinate Velocity Profile (BVP), extracted from CSIs into a hybrid deep learning model, which consists of a convolutional neural network (CNN) for spatial feature extraction and a recurrent neural network (RNN) for temporal modeling.

Semantic-Aware Networking: Utilizing semantic knowledge to enhance communication and networking performance has recently attracted significant interest [13], [14]. Most existing works focus on extracting human language-inspired semantic information to compress various forms of human generated signals, and improve communication efficiency and reliability [15]–[17]. For example, in [15] the authors adopted an attention mechanism-based solution to compress speech signals in which essential speech information is identified by providing higher weights to them when training the neural network. In [17], the authors considered a Transformer-based language text compression for maximizing the system capacity and minimizing the semantic errors by recovering the meaning of sentences. Multi-modal data compression was also investigated in [16], where a task-oriented semantic communications framework was proposed to unify the structure of transmitters for different tasks. In addition to compressing and recovering data bits, recent studies suggested that semantic information has a higher efficiency in recovering signals with high human-oriented perception quality. The so-called rate-distortion-perception tradeoff has been investigated in semantic communication [18], [19], where studies show that in some cases the receiver can directly infer the semantic information source satisfying certain distortion and perception constraints without requiring any data communication from the transmitter. Recently, semantic information has also been utilized to enable high-level reasoning and inference in communication networks [20], [21]. More specifically, the so-called implicit semantic-aware communication network was proposed in [20] in which the semantic correlations have been exploited to infer implicit information, such as clue information or background knowledge that are closely related to the data information sent over the network. In addition to communication networks, semantic knowledge has recently been extended to other fields, such as mmWave beam tracking [22], image and video segmentation [23],

emotional analysis [24], and affective computing [25]. In contrast to all these existing works, in this paper, we introduce the concept of physical-layer semantics to capture the impact of environmental and human-related features that influence the distribution of wireless sensing data. To the best of our knowledge, this is the first work that utilizes the semantic similarity of physical-layer features to transfer models between different locations and environmental scenarios.

Transfer Learning-based Wireless Sensing: To reduce the cost of model training, transfer learning methods have been recently applied to wireless sensing, with the goal to transfer knowledge obtained from a source domain to a target domain, so as to support a variety of wireless sensing tasks [26]. A straightforward idea is to extract domain-independent features from labeled samples in the source domain. For example, in [27]–[29], the authors show that adversarial architectures such as generative adversarial networks (GANs) can be used to learn the hidden relationships between the source inputs and the target outputs by combining a CNN feature extraction and a domain discriminator. Although integrating GANs into distributed wireless sensing solutions is a promising direction, it demands numerous ad-hoc "tricks" to achieve model convergence [30]. In [31], CrossSense was introduced as the state-of-the-art wireless transfer technique on WiFi-based gait identification and gesture recognition applications. To enable cross-domain sensing, CrossSense employs an artificial neural network (ANN) based mixture-of-experts strategy, where multiple specialized sensing models, or experts, are used to capture the mapping from diverse sourcing inputs to the targeting outputs.

Federated Learning-based Wireless Sensing: Federated learning (FL) is an emerging solution that enables distributed model training by utilizing model parameter sets instead of private data samples for sharing [32]. FL-based wireless sensing solutions have recently attracted significant interest due to their unique advantages, including decentralization, low communication overload, and privacy protection [26], [33]. For instance, the authors in [34] designed WiFederated for WiFi-based human activity recognition, which was the first FL-based wireless sensing framework proposed to overcome the challenge posed by the centralized model training paradigm. In [35] the authors introduced a cross-domain federated learning framework called CDFL, which aims at addressing the scarcity of labeled wireless data by generating simulated training data using a physical model guided by public datasets in other domains. Recent works [36], [37] also investigated distributed indoor localization by combining FL and wireless sensing based on receivers deployed across different locations.

3 SYSTEM MODEL AND PROBLEM FORMULATION

3.1 System Model

We consider human gesture recognition based on a distributed wireless sensing system consisting of one or more Wi-Fi transmitters and a set $\mathcal{K}$ of K receivers deployed at different locations across the considered area.

Each receiver records wireless signals (e.g., CSI data) that are reflected and scattered by human users when performing a set of gestures. We focus on the decentralized sensing scenario in which each receiver stores its recorded wireless signals locally which, due to the constraints in data privacy, cannot be exposed to others. We assume that only a subset of receivers $\mathcal{K}^L$ for $\mathcal{K}^L \subseteq \mathcal{K}$ can have labeled wireless sensing data samples. Each receiver in $\mathcal{K}^L$ can then construct a location-specific model to recognize different gestures of the human users based on its local dataset. There are some other receivers, denoted as subset $\mathcal{K}^N = \mathcal{K} \setminus \mathcal{K}^L$ that cannot have any labeled data and therefore cannot construct any local models using traditional supervised learning approaches. As mentioned earlier, due to the spatial heterogeneity of wireless sensing signals, receivers at different locations require different models to recognize the same gestures. In other words, receivers in $\mathcal{K}^N$ cannot directly utilize the gesture recognition models of receivers in $\mathcal{K}^L$ for their local gesture recognition tasks.

3.2 Physical-layer Semantics

We observe that the statistics of received CSI signals are closely related to the semantic information of the physical environment, such as the size and layout of rooms, the location of transmitters and receivers, and the human users' gesture profiles, such as the speed of movement of different body parts when performing different gestures, etc. Motivated by this observation, we investigate whether it is possible to develop a model transferring solution that allows one or a limited number of receivers with labeled data to transfer their locally trained models to other receivers, especially receivers without any labeled dataset, based on the correlations of environmental and gesture-related semantic features.

Let us first identify the key semantic features in wireless sensing systems that may influence the distribution of the CSI data recorded at each receiver. It is known that the CSI signal recorded by a receiver is mainly characterized by the wireless links connecting the transmitter and receiver, influenced by the gesture-performing human users as well as the physical objects located along side of the channels. More specifically, the CSI signal recorded by receiver k at arrival time α, subcarrier frequency θ, and antenna β can be written as [12]:

$$H_k(\alpha, \theta, \beta) = \left(\sum_{n \in \mathcal{L}_S} A_{k,n} e^{-j2\pi\theta\tau_{k,n}(\theta,\beta)} \right.$$
$$\left. + \sum_{m \in \mathcal{L}_M} A_{k,m}(\alpha) e^{-j2\pi\theta\tau_{k,m}(\alpha,\theta,\beta)} \right) e^{j\epsilon(\alpha,\theta,\beta)}, \quad (1)$$

where $\mathcal{L}_S$ and $\mathcal{L}_M$ are sets of stationary and dynamic path components, respectively, and $e^{j\epsilon(\alpha,\theta,\beta)}$ is the phase error caused by asynchronization between transceivers and hardware imperfection. For each propagation path l for $l \in \mathcal{L}_S \cup \mathcal{L}_M$, $A_{k,l}$ and $\tau_{k,l}$ are the channel attenuation factor and time delay, respectively. Here dynamic path components correspond to the received signals reflected by the moving targets, while the stationary path components

correspond to the signals received from the direct paths and the reflection signals from static objects such as walls and furniture. Since the CSI can only be sampled as discrete signals in time (packet), frequency (subcarrier), and space (antenna) [38], the time delay of static and dynamic signal paths, respectively, in (1) can be written as follows:

$$\tau_{k,n}(\theta, \beta) = \tau_{k,0} + \Delta\beta_{k,n} \cdot \varpi_{k,0}, \text{for } n \in \mathcal{L}_S \quad (2)$$
$$\tau_{k,m}(\alpha, \theta, \beta) = \tau_{k,0} - \frac{\rho_{k,0}}{\Delta\theta_{k,m}}\Delta\alpha_{k,m} + \Delta\beta_{k,m} \cdot \varpi_{k,0},$$
$$\text{for } m \in \mathcal{L}_M \quad (3)$$

where $\Delta\alpha_{k,l}$, $\Delta\theta_{k,l}$, $\Delta\beta_{k,l}$ for $l \in \mathcal{L}_S \cup \mathcal{L}_M$ are differences of packets, subcarriers, and spatial positions, respectively, between two consecutive CSI samples of $H_k(\alpha, \theta, \beta)$ in (1). $H_k(0, 0, 0)$ is defined as the CSI reference signal with the time delay $\tau_{k,0}$, DFS $\rho_{k,0}$ and AoA $\varpi_{k,0}$.

From (1), we can observe that the CSI signals recorded by receiver $k \in \mathcal{K}$ are closely related to the following two types of physical-layer semantics:

Environment-related semantics (E-semantics): include the semantic information related to the physical environment such as environmental layout and the relative locations and orientations of transmitters, receivers, and human users. We therefore can write the feature vector of E-semantics of receiver k as $\boldsymbol{u}_k = \langle A_{k,n}, \tau_{k,n}, \varpi_{k,n} \rangle_{n \in \mathcal{L}_S}$.

Gesture-related semantics (G-semantics): include the semantic information associated with gestures such as the users' body coordinates and movement patterns of gestures. We can write the feature vector of G-semantics of receiver k as $\boldsymbol{v}_k = \langle A_{k,m}, \tau_{k,m}, \varpi_{k,m}, \rho_{k,m} \rangle_{m \in \mathcal{L}_M}$.

We can then rewrite (1) into the following form:

$$H_k(\alpha, \theta, \beta) = \sum_{n \in \mathcal{L}_S} p_{k,n}(\theta, \beta; \boldsymbol{u}_k)$$
$$+ \sum_{m \in \mathcal{L}_M} q_{k,m}(\alpha, \theta, \beta; \boldsymbol{v}_k), \quad (4)$$

where $p_{k,n}(\theta, \beta; \boldsymbol{u}_k) = A_{k,n}e^{-j2\pi\theta\tau_{k,n}(\theta,\beta)+j\epsilon(\alpha,\theta,\beta)}$ and $q_{k,m}(\alpha, \theta, \beta; \boldsymbol{v}_k) = A_{k,m}(\alpha)e^{-j2\pi\theta\tau_{k,m}(\alpha,\theta,\beta)+j\epsilon(\alpha,\theta,\beta)}$ are stationary and dynamic path component signals, respectively.

We combine both E- and G-semantics and write the physical-layer semantic feature vector of wireless signals recorded by receiver k as $\boldsymbol{\phi}_k = \langle \boldsymbol{u}_k, \boldsymbol{v}_k \rangle$. We can observe that the physical-layer semantics are location-specific and therefore each receiver k has a unique semantic feature vector $\boldsymbol{\phi}_k$ which plays a key role in determining the probability distribution of the locally received CSI signals.

3.3 Physical-Layer Semantic-Aware Network

Let us now formally introduce the concept of physical-layer semantic-aware network (pSAN) as follows:

Definition 1. A *physical-layer semantic-aware network* (pSAN) is a wireless sensing network in which the physical-layer semantics, including both E- and G-semantics, can be aware, known, and utilized, by each receiver.

In pSAN, the similarity of physical-layer semantics between different receivers can be used to infer correlations between different location-specific models trained by these receivers. Recall that only a subset $\mathcal{K}^L$ of K^L receivers can

have labeled CSI signals. To simplify our description, we use k' for $k' \in \mathcal{K}^L$ to denote the kth receiver with labeled CSI data. Let $\mathcal{D}_{k'}$ be the set of labeled CSI data at receiver k'. We assume the labeled data samples at different receivers in $\mathcal{K}^L$ are associated with the same set of gesture classes. Similarly, let k'' for $k'' \in \mathcal{K}^N$ be the k''th receiver that does not have any labeled data.

The key idea is to establish a mapping function that converts different high-dimensional physical-layer semantics into the same low-dimensional semantic space to capture the similarity between the key statistic features of physical-layer semantics that determine the gesture recognition models trained by different receivers. Specifically, let $\bar{\phi}_k$ be the low-dimensional semantic vectors converted from ϕ_k to the semantic space for $k \in \mathcal{K}$. Common metrics for measuring semantic similarity include energy-based and statistic-based metrics. In this paper, we mainly focus on energy-based semantic similarity. We will present a formal definition and give a more detailed discussion in Section 6. Without loss of generality, in this paper, we use $S\left(\bar{\phi}_j, \bar{\phi}_k\right)$ to denote the semantic similarity between two semantic features ϕ_j and ϕ_k.

3.4 Problem Formulation

Each labeled CSI data $\zeta_{k',i} = \langle x_{k',i}, y_{k',i} \rangle$ recorded by receiver k' for $k' \in \mathcal{K}^L$ consists of a CSI signal $x_{k',i}$, e.g., an instance of CSI signal recorded by receiver k', and a class label $y_{k',i}$ that belongs to one of a set of gesture classes $\mathcal{Y}$. Let $\mathcal{D}_{k'}$ be the set of local training data samples at receiver k'. Each receiver $k' \in \mathcal{K}^L$ can then construct a local model $\boldsymbol{\omega}_{k'}$ by minimizing its local objective function,

$$\min_{\boldsymbol{\omega}_{k'}} F_{k'}\left(\boldsymbol{\omega}_{k'}\right) = \frac{1}{|\mathcal{D}_{k'}|} \sum_{\zeta_{k',i} \in \mathcal{D}_{k'}} \left[f_{k'}\left(\boldsymbol{\omega}_{k'}; \zeta_{k',i}\right) \right], \quad (5)$$

where $\boldsymbol{\omega}_{k'}$ is the model parameters of receiver k'.

We also need to learn a semantic-aware model transfer function to transfer models learned by receivers with labeled data to those receivers without any labeled data according to their semantic similarity. In our considered decentralized wireless sensing scenario, the CSI data recorded by each receiver cannot be exposed to others. It is however possible for the receivers to expose their locally trained models to other receivers. In the rest of this paper, we will develop a pSAN-based model aggregation and transfer approach in which each receiver $k'' \in \mathcal{K}^N$ can directly obtain a location-specific model by aggregating models that are already trained by receivers in $\mathcal{K}^L$.

The main objective is to design an appropriate model transfer approach, so the transferred model at receiver k'' can approach the locally optimal model $\boldsymbol{\omega}_{k''}^*$, i.e., we write the problem as follows:

$$\min_{\boldsymbol{\omega}_{k''}} \left[F_{k''}\left(\boldsymbol{\omega}_{k''}\right) - F_{k''}\left(\boldsymbol{\omega}_{k''}^*\right) \right], \quad \forall k'' \in \mathcal{K}^N, \quad (6)$$

where $\boldsymbol{\omega}_{k''}$ is the transferred model obtained by receiver k'' which, if we consider a linear model transfer framework, can be obtained as follows:

$$\boldsymbol{\omega}_{k''} = \sum_{k' \in \mathcal{K}'} \xi\left(S(\bar{\phi}_{k'}, \bar{\phi}_{k''})\right) \boldsymbol{\omega}_{k'}, \quad (7)$$

where $S(\bar{\phi}_{k'}, \bar{\phi}_{k''})$ denotes the semantic similarity between semantics $\bar{\phi}_{k'}$ of receiver k' and semantics $\bar{\phi}_{k''}$ of receiver k'', $\xi(\cdot)$ is a semantic-aware model transfer function that maps the semantic similarity between receivers k' and k'' to a normalized model aggregation coefficient value. We will give a more detailed discussion on how to obtain $S(\cdot, \cdot)$ and $\xi(\cdot)$ in Section 6 and prove the convergence result of our proposed solutions later in Section 7.
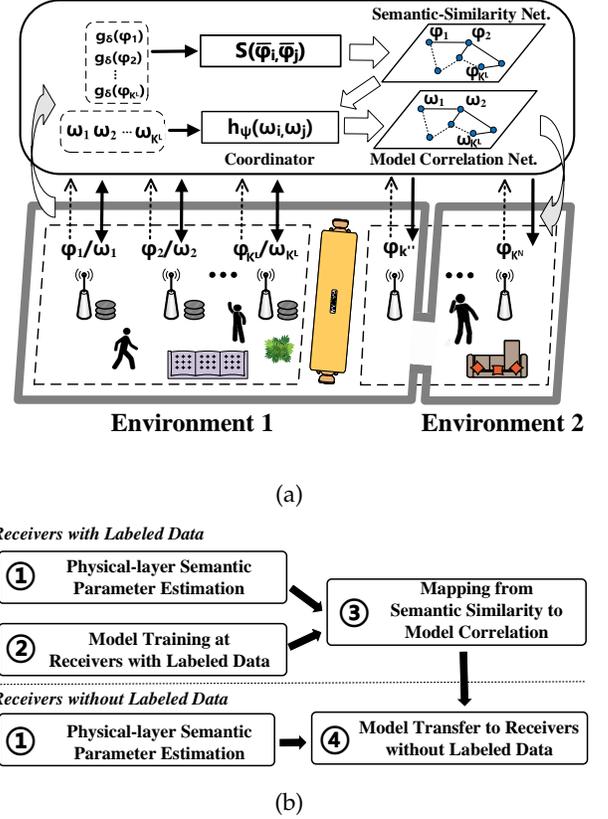

(a)

(b)

Fig. 2: (a) SANSee framework and (b) key training procedures.

4 SANSee Overview

The architectural framework and key training procedures of SANSee are illustrated in Fig. 2a and 2b, respectively. The detailed operations are described as follows:

Physical-layer Semantics Estimation: Each receiver needs to first estimate key physical-layer semantic parameters that influence its local CSI data. Note that estimating semantic parameters does not require any labeled CSI data. To estimate E- and G- semantics separately, each receiver needs to first separate its CSI signals by applying the high-pass and low-pass filters, respectively, and then apply the maximum likelihood estimation (MLE) approach to estimate the combination of different semantic parameters.

Mapping from Semantic Similarity to Model Correlation: After each receiver has successfully estimated its physical-layer semantics, we then need to construct a mapping function that can convert the semantic similarity to the model correlation between different receivers. To characterize the semantic similarity between different receivers, we introduce a low-dimensional semantic space

in which the distance between any two physical-layer semantics is proportional to their semantic similarity. We then construct a mapping function to map the high-dimensional semantic feature vector into the semantic space. We also introduce a correlation coefficient to characterize the model correlation between local models trained by different receivers. Finally, we design a novel loss function to simultaneously optimize parameters of the semantic mapping function and the calculation function of the model correlation coefficient to match semantic similarity with model correlations.

Model Training at Receivers with Labeled Data: All the receivers with labeled data will jointly construct their location-specific models. We adopt a personalized federated learning-based solution for receivers to collaboratively train their location-specific models without exposing their local datasets. After successfully training their models, all the receivers with labeled data will link their models with their physical-layer semantics and establish a mapping function to convert semantic similarity to model correlation coefficients.

Model Transfer at Receivers without Labeled Data: Each receiver without labeled data will rely on the coordinator to construct its location-specific model based on the correlated model trained by receivers with labeled data. More specifically, each receiver without labeled data will submit its locally estimated semantic features to the coordinator. The coordinator will then apply the previously constructed mapping function to calculate the model correlation coefficients for all the correlated models obtained by receivers with labeled data, and finally send the aggregated model to each corresponding receiver.

5 PHYSICAL-LAYER SEMANTICS ESTIMATION

The first step in pSAN is to quantify the impact of physical-layer semantics on the CSI data recorded by each receiver. From (1), we can observe that, the raw CSI signal $H_k(\alpha, \theta, \beta)$ obtained by each receiver consists of phase error term $e^{j\epsilon(\alpha,\theta,\beta)}$ which may result in inaccurate estimation of physical-layer semantics. This issue can be addressed when the receiver has two or more antennas, in which the phase error term can be canceled by performing conjugate multiplication and amplitude adjustment on CSI signals received by two antennas [11]. Let $\hat{H}_k(\alpha, \theta, \beta)$ be the phase error-canceled version of the CSI signal of receiver k. We also use $\hat{q}_{k,m}$ and $\hat{p}_{k,n}$ to denote dynamic and stationary path components in $\hat{H}_k(\alpha, \theta, \beta)$, respectively.

By applying a high-pass filter, we can separate the sum of dynamic components related to G-semantics $\boldsymbol{v}_k$ from the raw CSI of receiver $k \in \mathcal{K}$, denoted as $\hat{H}_k^M(\alpha, \theta, \beta) = \sum_{m \in \mathcal{L}_M} \hat{q}_{k,m}(\alpha, \theta, \beta; \boldsymbol{v}_k)$. We can then adopt the maximum likelihood estimation (MLE) to estimate the G-semantics parameters consisting of a collection of parameters of all dynamic signal components, i.e., $\boldsymbol{v}_k = \{\boldsymbol{v}_{k,m}\}_{m \in \mathcal{L}_M}$ with $\boldsymbol{v}_{k,m} = \langle A_{k,m}, \tau_{k,m}, \varpi_{k,m}, \rho_{k,m}\rangle$. More specifically, the G-semantics $\boldsymbol{v}_k^*$ of receiver k can be estimated by solving

the following problem:

$$\boldsymbol{v}_k^* = \arg\max_{\boldsymbol{v}_k}\{- \sum_{\alpha \in \mathcal{A}, \theta \in \boldsymbol{\Theta}, \beta \in \mathcal{B}} |\hat{H}_k^M(\alpha, \theta, \beta)$$
$$- \sum_{m \in \mathcal{L}_M} q_{k,m}(\alpha, \theta, \beta; \boldsymbol{v}_{k,m})|^2\}, \quad (8)$$

where $\hat{H}_k^M(\alpha, \theta, \beta)$ is the obtained from real-measured CSI signal and $q_{k,m}(\alpha, \theta, \beta; \boldsymbol{v}_{k,m})$ is the estimated components. $\mathcal{A}$, $\boldsymbol{\Theta}$, $\mathcal{B}$ are the sets of possible packets, subcarriers, and antennas, i.e., $\alpha \in \mathcal{A}, \theta \in \boldsymbol{\Theta}, \beta \in \mathcal{B}$.

Similarly, we can extract the sum of stationary components $\hat{H}_k^S(\theta, \beta) = \sum_{n \in \mathcal{L}_S} p_{k,n}(\theta, \beta; \boldsymbol{u}_k)$ related to E-semantics by applying a low-pass filter and estimate parameters in E-semantics, i.e., $\boldsymbol{u}_k = \{\boldsymbol{u}_{k,n}\}_{n \in \mathcal{L}_S}$ with $\boldsymbol{u}_{k,n} = \langle A_{k,n}, \tau_{k,n}, \varpi_{k,n}\rangle$ as follows:

$$\boldsymbol{u}_k^* = \arg\max_{\boldsymbol{u}_k}\{- \sum_{\theta \in \boldsymbol{\Theta}, \beta \in \mathcal{B}} |\hat{H}_k^S(\theta, \beta)$$
$$- \sum_{n \in \mathcal{L}_S} p_{k,n}(\theta, \beta; \boldsymbol{u}_{k,n})|^2\}, \quad (9)$$

where $\hat{H}_k^S(\alpha, \theta, \beta)$ is obtained from the real-measured CSI signal and $p_{k,n}(\theta, \beta; \boldsymbol{u}_{k,n})$ is the estimated components.

We can observe that, it is generally difficult to derive closed-form solutions of $\boldsymbol{v}_k^*$ and $\boldsymbol{u}_k^*$ in (8) and (9). We can however adopt a modified Space Alternating Generalized Expectation Maximization (mSAGE) algorithm to estimate the values of $\boldsymbol{v}_k^*$ and $\boldsymbol{u}_k^*$ using an iteration-based approach [39]. We use superscript t to denote the operation in the tth iteration. The mth dynamic signal path component can be calculated by first performing the expectation step as follows:

$$q_{k,m}^{t+1}(\alpha, \theta, \beta; \boldsymbol{v}_{k,m}^t) = q_{k,m}^t(\alpha, \theta, \beta; \boldsymbol{v}_{k,m}^t) \quad (10)$$
$$+ \pi_H \left(\hat{H}_k^M(\alpha, \theta, \beta) - \sum_{m' \in \mathcal{L}_M} q_{k,m}^t(\alpha, \theta, \beta; \boldsymbol{v}_{k,m'}^t)\right),$$

where $\boldsymbol{v}_{k,m'}^t$ is the G-semantics of the m-th path estimated in the tth iteration of receiver k, and π_H is the non-negative step size and its default value can be set as 1. We then obtain the optimal value of parameter $\boldsymbol{v}_{k,m}^*$ by maximizing the magnitude of the signal received at the mth signal path component $z_{k,m}(\tau, \varpi, \rho; q_{k,m}^{t+1}) = \sum_{\alpha \in \mathcal{A}, \theta \in \boldsymbol{\Theta}, \beta \in \mathcal{B}} |e^{2\pi(\Delta\theta_{k,m}\tau_{k,m} + f_c\Delta\beta_{k,m}\varpi_{k,m} - \Delta\alpha_{k,m}\rho_{k,m})} q_{k,m}^{t+1}(\alpha, \theta, \beta; \boldsymbol{v}_{k,m}^t)|^2$, i.e. $\boldsymbol{v}_{k,m}^*$ is given by,

$$\boldsymbol{v}_{k,m}^* = \arg\max_{\boldsymbol{v}_{k,m}} z_{k,m}(\tau, \varpi, \rho; q_{k,m}^{t+1}), \quad (11)$$

where f_c is the carrier frequency of the wireless channel, and $\Delta\alpha_{k,m}$, $\Delta\theta_{k,m}$, $\Delta\beta_{k,m}$ are defined previously in (3). To solve (11), we apply the following steps to sequentially estimate each individual parameter $\tau_{k,m}^{t+1}$, $\varpi_{k,m}^{t+1}$, $\rho_{k,m}^{t+1}$, and $A_{k,m}^{t+1}$ in $\boldsymbol{v}_{k,m}$ as follows:

$$\tau_{k,m}^{t+1} = \arg\max_{\tau} |z_{k,m}(\tau, \varpi_{k,m}^t, \rho_{k,m}^t; q_{k,m}^{t+1}(\alpha, \theta, \beta; \boldsymbol{v}_{k,m}^t))|^2, \quad (12)$$

$$\varpi_{k,m}^{t+1} = \arg\max_{\varpi} |z_{k,m}(\tau_{k,m}^{t+1}, \varpi, \rho_{k,m}^t; q_{k,m}^{t+1}(\alpha, \theta, \beta; \boldsymbol{v}_{k,m}^t))|^2, \quad (13)$$

$$\rho_{k,m}^{t+1} = \arg\max_{\rho} |z_{k,m}(\tau_{k,m}^{t+1}, \varpi_{k,m}^{t+1}, \rho; q_{k,m}^{t+1}(\alpha, \theta, \beta; \boldsymbol{v}_{k,m}^t))|^2, \quad (14)$$

$$A_{k,m}^{t+1} = \frac{z_{k,m}(\tau_{k,m}^{t+1}, \varpi_{k,m}^{t+1}, \rho_{k,m}^{t+1}; q_{k,m}^{t+1}(\alpha, \theta, \beta; \boldsymbol{v}_{k,m}^t))}{|\mathcal{A}| \cdot |\boldsymbol{\Theta}| \cdot |\mathcal{B}|}. \quad (15)$$

The above iteration process ends when the difference between two successive estimations of $\boldsymbol{v}_{k,m}$ is within a pre-defined threshold ς. For stationary component signal estimation, we can follow a similar approach to estimate the parameters of $\boldsymbol{u}_k$ for receiver k. The detailed procedures of the physical-layer semantics estimation process are summarized in Algorithm 1.

Algorithm 1 Physical-layer Semantics Estimation Algorithm of Receiver k

---

Input: CSI $H_k(\alpha, \theta, \beta)$; Numbers of estimated paths L_M and L_N; Pre-defined threshold ς; Initial iteration $t = 0$; Initial values $\boldsymbol{u}_k = 0, \boldsymbol{v}_k = 0$.
Output: Physical-layer semantics $\phi_k = \langle \boldsymbol{u}_k^*, \boldsymbol{v}_k^* \rangle$.

1:   Cancel $H_k(\alpha, \theta, \beta)$ by denoising and obtain $\hat{H}_k(\alpha, \theta, \beta)$;
2:   **While** $\|\boldsymbol{v}_{k,m}^t - \boldsymbol{v}_{k,m}^{t+1}\| \leq \varsigma$ **do**
3:       **For** $m = 1, \cdots, L_M$ **do**
4:           Apply a high-pass filter to obtain $\hat{H}_k^M(\alpha, \theta, \beta)$;
5:           Calculate $q_{k,m}^{t+1}(\alpha, \theta, \beta; \boldsymbol{v}_{k,m}^t)$ by using (10);
6:           Estimate parameters of $\boldsymbol{v}_{k,m}^{t+1}$ by using (12)-(15);
7:       **End for**
8:       $t = t + 1$;
9:   **End while**
10: **While** $\|\boldsymbol{u}_{k,n}^t - \boldsymbol{u}_{k,n}^{t+1}\| \leq \varsigma$ **do**
11:      **For** $n = 1, \cdots, L_N$ **do**
12:          Apply a low-pass filter to obtain $\hat{H}_k^S(\theta, \beta)$;
13:          Calculate $p_{k,n}^{t+1}(\theta, \beta; \boldsymbol{u}_{k,n}^t)$ by substituting $\hat{H}_k^S$ into (10);
14:          Estimate $\boldsymbol{u}_{k,n}^{t+1}$ by substituting $p_{k,n}^{t+1}$ into (12)-(15);
15:      **End for**
16:      $t = t + 1$;
17: **End while**

---

6 MODEL TRAINING AND TRANSFER

6.1 Semantic Similarity and Model Correlations

From the previous discussion, we can observe that the physical-layer semantics directly affect the distributions of the CSI data at each receiver. It is known that, for a given algorithmic framework, the distribution of training dataset and the resulting model are in one-to-one correspondence. Thus, in this section, we aim to develop a mapping function that converts the semantic similarity to the correlations of models.

Motivated by the fact that physical-layer semantics of each receiver consist of multiple key parameters that have different impacts on the performance of different gesture-recognition tasks, we need to first convert the high-dimensional physical-layer semantics of different receivers into a low-dimensional space referred to as the (physical-layer) semantic space. In the semantic space, the distance between different semantics of different receivers is proportional to the correlations of their local gesture recognition models, e.g., the larger the distance (similarity) between receivers' semantics, the higher the correlations between different local models of different receivers. In this way, we can use the semantic similarity to transfer models from some receivers, e.g., receivers with labelled data, to other receivers, e.g., receivers without labeled data, without requiring any extra model training.

In this paper, we consider a neural network-based mapping function to convert the high-dimensional physical-layer semantics ϕ_k into the low-dimensional version $\bar{\phi}_k$ in the semantic space. We can write the mapping function that outputs the low-dimensional semantic representation as $\bar{\phi}_k = g_\delta(\phi_k)$, where δ is the parameters of the mapping function.

Let $S(\bar{\phi}_k, \bar{\phi}_j)$ be the semantic similarity between receivers k and j in the semantic space. We consider a general framework in which semantic similarity can be measured using different metrics. For example, if the Euclidean distance has been adopted to measure similarity between two semantics $\bar{\phi}_k$ and $\bar{\phi}_j$ in semantic space, we can write:

$$S(\bar{\phi}_k, \bar{\phi}_j) = S(g_\delta(\phi_k), g_\delta(\phi_j)) = |\bar{\phi}_k - \bar{\phi}_j|^2. \quad (16)$$

We can also use other types of metrics such as statistic-based similarity metrics, including cross-entropy (CE) and Jensen–Shannon divergence (JSD), by following the same line in [20].

Next, we need to define the correlation between gesture-recognition models trained based on datasets available at different receivers. In this paper, we adopt a linear correlation in which the correlation between different models $\boldsymbol{\omega}_j$ and $\boldsymbol{\omega}_k$ is characterized by a linear coefficient $\xi_{j,k}$. If suppose model $\boldsymbol{\omega}_j$ is correlated with a set of models, e.g., $\{\boldsymbol{\omega}_k\}_{k \in \mathcal{K}^L}$ for $j \notin \mathcal{K}^L$, we then can write model $\boldsymbol{\omega}_j$ as a linear combination of all the correlated models with normalized coefficients given by $\boldsymbol{\omega}_j = \sum_{k \in \mathcal{K}^L} \xi_{j,k} \boldsymbol{\omega}_k$, where $\xi_{j,k}$ satisfies $0 \leq \xi_{j,k} \leq 1$ and $\sum_{k \in \mathcal{K}^L} \xi_{j,k} = 1$. Suppose the model correlation coefficient $\xi_{j,k}$ can also be learned by a neural network h_ψ with parameter ψ, i.e., we can write $\xi_{j,k} = h_\psi(\boldsymbol{\omega}_j, \boldsymbol{\omega}_k)$.

Finally, we can use the following loss function to train parameters δ and ψ to match the semantic similarity with the model correlation:

$$\mathcal{L}(\delta, \psi) = \sum_{j,k \in \mathcal{K}^L} |h_\psi(\boldsymbol{\omega}_j, \boldsymbol{\omega}_k) - S(\bar{\phi}_k, \bar{\phi}_j)|^2. \quad (17)$$

The models δ and ψ can be trained at the same time by minimizing the above loss function using the standard SGD approach.

In SANSee, δ and ψ are first trained based on the set of receivers with labeled data $\mathcal{K}^L$. The receivers without labeled data in $\mathcal{K}^N$ can then directly obtain their local models by performing a linear combination operation on the set of models $\{\boldsymbol{\omega}_k\}_{k \in \mathcal{K}^L}$. We will give a more detailed discussion on the model construction process at receivers in $\mathcal{K}^L$ as well as the model transfer process from receivers in $\mathcal{K}^L$ to receivers $\mathcal{K}^N$ in the next section.

6.2 Model Training at Receivers with labelled data

In this paper, we follow a commonly adopted FL setting in which receivers optimize their model parameters to minimize the loss functions based on their local data distributions. In other words, for a given model, the optimal model parameters obtained based on the local data minimize the loss function and maximize the output accuracy of the trained model. The optimal parameters of the trained models directly reflect the correlation between the data distributions of different receivers and therefore can be used to decide the set of receivers with similar data distributions. The above results have been verified both theoretically and practically in many FL-based applications

and have already served as the foundation of many well-developed personalized FL solutions [40]–[45].

In fact, the difference between model parameters learned by different receivers due to different distributions of the local datasets is commonly referred to as the client drift problem. This problem results in slow convergence and even divergence of the model training process in model-aggregation-based FL approaches [46]. To address the client drift problem, an attention-inducing function $\lambda \sum_{k' < j'} R\left(\|\boldsymbol{\omega}_{k'} - \boldsymbol{\omega}_{j'}\|^2\right)$ is introduced in the regularized loss function, which improves the collaboration between the personalized models trained by different receivers. The attention-inducing function enhances the convergence and performance of personalized models through an attentive message-passing mechanism, which is model agnostic and can coordinate various intermediate results, with proven convergence for both convex and non-convex models [47]. Specifically, we consider an attention-inducing function-based personalized federated learning solution in which all receivers in $\mathcal{K}^L$ collaborate in training a set of $\mathcal{K}^L$ location-specific models, denoted as $\boldsymbol{\Omega} = \langle \boldsymbol{\omega}_k \rangle_{k \in \mathcal{K}^L}$ by minimizing the following objective functions:

$$\mathcal{J}(\boldsymbol{\Omega}) := \sum_{k' \in \mathcal{K}^L} \left(F_{k'}(\boldsymbol{\omega}_{k'}) + \lambda \sum_{k' < j'} R\left(\|\boldsymbol{\omega}_{k'} - \boldsymbol{\omega}_{j'}\|^2\right) \right), \quad (18)$$

where $\lambda > 0$ is a non-negative collaboration parameter, $R(\|\boldsymbol{\omega}_{k'} - \boldsymbol{\omega}_{j'}\|^2)$ is a regularizer which is an attention-inducing function included here to encourage collaborations between receivers with correlated models. In particular, we follow a commonly adopted setting [47] and use the negative exponential function to characterize the difference between models $\boldsymbol{\omega}_{k'}$ and $\boldsymbol{\omega}_{j'}$, defined as follows:

$$R(\|\boldsymbol{\omega}_{k'} - \boldsymbol{\omega}_{j'}\|^2) = 1 - e^{-\|\boldsymbol{\omega}_{k'} - \boldsymbol{\omega}_{j'}\|^2/\sigma_R}, \quad (19)$$

where σ_R is the difference parameter that controls the relative difference between models. The added attention-inducing function in the objective function in (18) is an increasing function of the difference between model parameters of receivers. Thus, when minimizing the objective function at a receiver, other models learned by receivers with higher (lower) similarity in the local data distributions will have higher (lower) weights. Moreover, the regularization also smooths the difference between the model parameters at different receivers. This further reduces the variations of the model parameter differences, especially at the beginning of the model training process, which further improves the convergence and robustness of the personalized model aggregation. We then describe the detailed personalized model training process.

In this paper, we adopt a standard SGD-based FL setting as introduced in [33] to iteratively construct personalized models for receivers in the set $\mathcal{K}^L$. Specifically, a coordinator is pre-assigned and announced to all the receivers, which would periodically upload their local model parameters to the coordinator for model aggregation and download the updated models for the next round of local model training. The proposed model is flexible; the coordinator is a logical entity deployed at any receiver, e.g., a coordination receiver, or a physical entity installed at a dedicated central server. In the former case,

all other receivers periodically upload their intermediate local models to the coordinating receiver, which in turn aggregates the received models with its own model. In the latter case, all receivers upload their intermediate local models to the central server for model aggregation once in a while. Both scenarios have already been widely applied in many FL applications. In the rest of this section, we use the superscript $(\cdot)^{t,e}$ to denote the parameters in eth local iteration of the tth global coordination round, i.e., $\boldsymbol{\omega}^{t,e}$ is the model downloaded from the coordinator at the beginning of the tth round. In the tth coordination round, each receiver $k' \in \mathcal{K}^L$ updates its local model as follows:

$$\boldsymbol{\omega}_{k'}^{t,e+1} = \boldsymbol{\omega}_{k'}^{t,e} - \eta \nabla \widetilde{F}_{k'}(\boldsymbol{\omega}_{k'}^{t,e}), \text{ for } e = 0, \ldots, E-1 \quad (20)$$

where $\boldsymbol{\omega}_{k'}^{t,e}$ denotes the local model of receiver k' in the eth iteration in the tth coordination round, η is the local learning rate, and $\nabla \widetilde{F}_{k'}(\boldsymbol{\omega}_{k'}^{t,e})$ is the unbiased stochastic gradient. At the end of the Eth local iteration, receivers will upload models $\{\boldsymbol{\omega}_1^{t,E}, \cdots, \boldsymbol{\omega}_{K^L}^{t,E}\}$ to the coordinator for global model updating. At the coordinator, the following step will be performed for each receiver k' to obtain the next-round model $\boldsymbol{\omega}_{k'}^{t+1}$ for each receiver k' as follows, for $k', j' \in \mathcal{K}^L$:

$$\boldsymbol{\omega}_{k'}^{t+1} = \boldsymbol{\omega}_{k'}^{t,E} - \sum_{k' \neq j'} \widetilde{\eta} \lambda \nabla R(\|\boldsymbol{\omega}_{k'}^{t,E} - \boldsymbol{\omega}_{j'}^{t,E}\|^2), \quad (21)$$

where $\widetilde{\eta} = \eta E$ is the step size. Repeat the above processes until the preset target loss $\epsilon_{\mathcal{J}}$ is reached.

In fact, the step in (21) at the coordinator is in essence to update the model for each receiver k' by performing a linear combination given by

$$\boldsymbol{\omega}_{k'}^{t+1} = \sum_{j' \in \mathcal{K}^L} \xi_{k',j'}^t \boldsymbol{\omega}_{j'}^{t,E} \quad (22)$$

where $\xi_{k',j'}^t$ is given by

$$\xi_{k',j'}^t = \begin{cases} \widetilde{\eta} \lambda R^{'}(\|\boldsymbol{\omega}_{k'}^{t,E} - \boldsymbol{\omega}_{j'}^{t,E}\|^2), & k' \neq j', \\ 1 - \widetilde{\eta} \lambda \sum_{j' \neq k'}^{K^L} R^{'}(\|\boldsymbol{\omega}_{k'}^{t,E} - \boldsymbol{\omega}_{j'}^{t,E}\|^2), & k' = j', \end{cases} \quad (23)$$

where $R^{'}(\|\boldsymbol{\omega}_{k'}^{t,E} - \boldsymbol{\omega}_{j'}^{t,E}\|^2) = \frac{e^{-\|\boldsymbol{\omega}_{k'}^{t,E} - \boldsymbol{\omega}_{j'}^{t,E}\|^2/\sigma_R}}{\sigma_R}$. Note that, the value of $R(\|\boldsymbol{\omega}_{k'}^{t,E} - \boldsymbol{\omega}_{j'}^{t,E}\|^2)$ decreases as the model correlation between models $\boldsymbol{\omega}_{k'}^{t,E}$ and $\boldsymbol{\omega}_{j'}^{t,E}$ increase. Also, since $0 \leq R^{'}(\|\boldsymbol{\omega}_{k'}^{t,E} - \boldsymbol{\omega}_{j'}^{t,E}\|^2) \leq \frac{1}{\sigma_R}$, we can ensure $0 \leq \xi_{k',j'}^t \leq 1$ by choosing a proper local learning rate, e.g., $\eta \leq \frac{1}{\lambda E (K^L - 1)}$. The coordinator needs to perform only simple linear combining operations based on the models uploaded by a limited number of receivers with labeled data. Thus its computational load is negligible compared to that of the local model training process at each receiver. More specifically, in SANSee, we follow similar model aggregation operations as the existing personalized FL solutions in which the coordinator performs linear combining of model parameters received from the receivers for personalized model coordination. The overhead of such a model aggregation approach is generally considered negligible by many existing works in FL [48].

Algorithm 2 Model Training Algorithm

Input: Target loss ϵ_J; Local SGD steps E; Set of receivers with labelled data $\mathcal{K}^L$; labelled data $\{\mathcal{D}_1, \ldots, \mathcal{D}_{K^L}\}$;
Output: Personalized models of labeled receivers $\{\boldsymbol{\omega}_0^T, \ldots, \boldsymbol{\omega}_{K^L-1}^T\}$.

1:     Server broadcasts an initial model $\boldsymbol{\omega}_0$ to all receivers in $\mathcal{K}^L$;
2:     **While** $\mathcal{J}(\boldsymbol{\Omega}) \geq \epsilon_J$ **do**
3:       **For** receiver $k' \in \mathcal{K}^L$ **in parallel do**
4:         **For** $e = 0, \cdots, E - 1$ **do**
5:           Uniformly sample a mini-batch $\zeta_{k'}^{t,e}$ from $\mathcal{D}_{k'}$;
6:           Perform SGD iterations on $\boldsymbol{\omega}_{k'}^{t,e}$ by using (20);
7:         **End for**
8:       **End parallel for**
9:       **For** $k' \in \mathcal{K}^L$ **do on coordinator**
10:        Obtain coefficient $\xi_{k',j'}^t$ by using (23);
11:        Update next-round model $\boldsymbol{\omega}_k^{t+1}$ by using (22);
12:       **End for on coordinator**
13:    **End for**

6.3 Model Transfer to Receivers without labelled data

Let us now develop a model transfer solution that maps the personalized models constructed by receivers with labelled data to receivers without any labelled data. Specifically, each receiver $k' \in \mathcal{K}^L$ with a labeled dataset first establishes a semantic mapping pair $\langle \boldsymbol{\phi}_{k'}, \boldsymbol{\omega}_{k'} \rangle$ consisting of its location-specific semantics $\boldsymbol{\phi}_{k'}$ obtained in Section 5 and its local model $\boldsymbol{\omega}_{k'}$ constructed in Section 6.2. We can then follow the same line as Section 6.1 to jointly develop two modules: a semantic mapping functional module $g_\delta(\cdot)$ with parameter δ and a model correlation functional module $h_\psi(\cdot, \cdot)$ with parameter ψ.

The detailed procedures for implementing model transfer in SANSee are illustrated in Fig. 3. The semantic mapping functional module $g_\delta(\cdot)$ with parameter δ is implemented based on a 4-convolutional block-based CNN architecture in which each block consists of a 3×3 convolutional layer followed by a batch normalization and a ReLU layer. Two max-pool layers are then inserted after the first two blocks to extract important features while simultaneously reducing the data dimensions. After that, the resulting low-dimensional semantics $\{\bar{\boldsymbol{\phi}}_1, \cdots, \bar{\boldsymbol{\phi}}_{k'}\}$ are concatenated and fed into a feature concatenation layer, followed by two convolutional blocks, a fully connected ReLU layer and a fully connected sigmoid layer that outputs the semantic similarity between any pairs of input physical-layer semantics. The model correlation functional module $h_\psi(\cdot, \cdot)$ with parameter ψ is implemented using the convolutional block concatenated with two fully connected layers. Finally, the objective loss function $\mathcal{L}(\delta, \psi)$ given in (17) is used to establish the mapping relationship between semantic similarity and model correlations. To minimize the loss function $\mathcal{L}(\delta, \psi)$, we jointly optimize both functional modules by solving the following problem:

$$\langle \delta^*, \psi^* \rangle = \arg\min_{\langle \delta, \psi \rangle} \mathcal{L}(\delta, \psi). \tag{24}$$

In this case, receivers with no any labeled data can obtain a location-specific model by performing linear combinations of all personalized models at receivers with labeled data, i.e., the location-specific model $\boldsymbol{\omega}_{k''}$ of receiver $k'' \in \mathcal{K}^N$ can be calculated as $\boldsymbol{\omega}_{k''} = \sum_{k' \in \mathcal{K}'} \xi_{k'',k'} \boldsymbol{\omega}_{k'}$, where $\xi_{k'',k'}$ is the model aggregation coefficient predicted by the optimized semantic mapping functional module, i.e.,
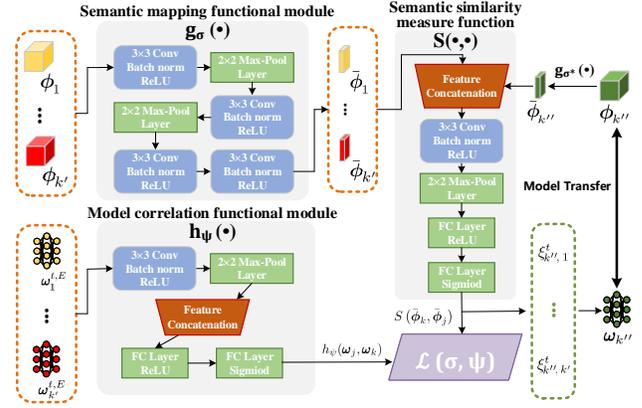

Fig. 3: Detailed architectural components of the proposed model transfer solution.

$\xi_{k'',k'} = S(g_{\delta^*}(\boldsymbol{\phi}_{k''}), g_{\delta^*}(\boldsymbol{\phi}_{k'}))$. SANSee does not require any labelled data at the target receivers. Furthermore, the model transfer process involves only linear operations summation and therefore, compared to existing transfer learning solutions [29]–[31]. SANSee significantly reduces the data labelling overhead as well as the required computational cost at the target receivers. We illustrate the detailed procedures of model transfer in Algorithm 3. As will be proved in the next section, the model obtained by each receiver $k'' \in \mathcal{K}^N$ without labeled data can approach to the real local model $\boldsymbol{\omega}_{k''}^*$.

Algorithm 3 pSAN-based Model Transfer Algorithm

Input: Raw CSI samples of all receivers.
Output: Transfer models $\{\boldsymbol{\omega}_{k''}\}_{k'' \in \mathcal{K}^N}$ of receivers in $\mathcal{K}^N$.

1:     Estimate physical-layer semantics $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\}$ of all receivers by using Alg. 1;
2:     Obtain local models $\{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_{K^L}\}$ of receivers in $\mathcal{K}^L$ by using Alg. 2;
3:     Construct a set of semantic mapping pairs $\{\boldsymbol{\phi}_{k'}, \boldsymbol{\omega}_{k'}\}_{k' \in \mathcal{K}^L}$;
4:     Train the two functional modules by minimizing (17) ;
5:     **For** $k'' \in \mathcal{K}^N$ **do on coordinator**
6:       Calculate aggregation coefficients $\{\xi_{k'',1}, \cdots, \xi_{k'',K^L}\}$ by using the optimized modules;
7:       Obtain the transfer model $\boldsymbol{\omega}_{k''}$ by performing a linear combination;
8:     **End for on coordinator**

7 THEORETICAL RESULTS

In this section, we present the theoretical results related to our proposed SANSee architecture. As mentioned earlier, SANSee is a distributed personalized model construction framework that involves two major steps: (1) Local model training: it first trains a set of models at the receivers with labelled data and, (2) model transfer: these trained models will be transferred to new receivers at novel locations without requiring any labelled data. In the rest of this section, we derive theoretical bounds of the following two types of errors:

(1)   **(Local) Model Training Error**: corresponds to the performance gap between the models $\boldsymbol{\Omega} = \langle \boldsymbol{\omega}_k \rangle_{k \in \mathcal{K}^L}$ trained by receivers based on their locally recorded datasets and the ground truth

model $\boldsymbol{\Omega}^* = \langle\boldsymbol{\omega}_k^*\rangle_{k\in\mathcal{K}^L}$, given by $\mathbb{E}[\mathcal{J}(\boldsymbol{\Omega}) - \mathcal{J}(\boldsymbol{\Omega}^*)]$ where $\mathcal{J}(\boldsymbol{\Omega})$ is defined previously in (18). We will a present detailed discussion in Section 7.1.

(2) **Model Transfer Error**: corresponds to the error of the transferred models obtained by the receivers without labelled data and the ground truth model, defined previously in (6). We will present a detailed discussion in Section 7.2.

7.1 Model Training Error

We use superscript T to denote the models trained in the Tth coordination round, e.g., we use $\boldsymbol{\Omega}^0$ and $\boldsymbol{\Omega}^T$ to denote the models in the 0th (initial model vector) and Tth coordination round. We can then prove the following result about the model training error.

Theorem 1. Suppose the following assumptions hold:

Assumption 1: (Strong Convexity) $F_1, \cdots, F_K$ are all μ-convex: i.e., $\frac{\mu}{2}\|\boldsymbol{\nu} - \boldsymbol{\omega}\|^2 \leq F_k(\boldsymbol{\nu}) - F_k(\boldsymbol{\omega}) -\langle\nabla F_k(\boldsymbol{\omega}), \boldsymbol{\nu} - \boldsymbol{\omega}\rangle$, for all $\boldsymbol{\nu}, \boldsymbol{\omega} \in \mathbb{R}^d$ and $k \in \mathcal{K}$,

Assumption 2: (Lipschitz Smoothness) $F_1, \cdots, F_K$ are all L-smooth: i.e., $\frac{L}{2}\|\boldsymbol{\nu} - \boldsymbol{\omega}\|^2 \geq F_k(\boldsymbol{\nu}) - F_k(\boldsymbol{\omega}) -\langle\nabla F_k(\boldsymbol{\omega}), \boldsymbol{\nu} - \boldsymbol{\omega}\rangle$, for all $\boldsymbol{\nu}, \boldsymbol{\omega} \in \mathbb{R}^d$ and $k \in \mathcal{K}$,

Assumption 3: (Bounded Variance) The variance of stochastic gradients on all local objective functions is bounded: $\mathbb{E}\|\nabla\widetilde{F}_k(\boldsymbol{\omega}, \zeta_k) - \nabla F_k(\boldsymbol{\omega})\|^2 \leq \sigma_F^2$, for all $\boldsymbol{\omega} \in \mathbb{R}^d$ and $k \in \mathcal{K}$,

Assumption 4: (Bounded Gradient) The gradient of the attention-inducing function is bounded: $\nabla R(\|\boldsymbol{\nu} - \boldsymbol{\omega}\|^2) \leq \kappa_R$, for all $\boldsymbol{\nu}, \boldsymbol{\omega} \in \mathbb{R}^d$.

Then, there exist $\lambda > 2L$, $T \geq \frac{4}{\tilde{\eta}_1\mu}$, and learning rate $\eta \leq \frac{\tilde{\eta}_1}{E}$ such that

$$\mathbb{E}[\mathcal{J}(\boldsymbol{\Omega}^T) - \mathcal{J}(\boldsymbol{\Omega}^*)] \leq \mathcal{O}\left(\|\boldsymbol{\Omega}^0 - \boldsymbol{\Omega}^*\|^2 e^{\frac{-\tilde{\eta}_1\mu T}{4}} + \frac{(1 + \mu T)(E\Gamma^L + \sigma_F^2/B)}{\mu^3 T^2 E K^L}\right), \quad (25)$$

where $\tilde{\eta}_1 \coloneqq \left(12(L + \lambda\kappa_R) + \frac{128\lambda\kappa_R L^2}{\mu^2} + \frac{96L^2}{\mu}\right)^{-1}$, $\Gamma^L = \sum_{k'\in\mathcal{K}^L}\|\nabla F_{k'}(\boldsymbol{\omega}_{k'}^*)\|^2$, and B is mini-batch size. $\mathcal{O}(\cdot)$ is the big-O notation which ignores poly-logarithmic and constant numerical factors.

Proof: See Appendix A. $\qquad\square$

We can observe that the assumptions introduced in the above theorem are reasonable in many practical scenarios. More specifically, as discussed in [42], [46], Assumptions 1-3 are satisfied by many commonly adopted loss functions such as cross-entropy, L2 regularization, etc. Assumption 4 can also be achieved by choosing many commonly used regularizers such as the negative exponential function [40].

We can observe from Theorem 1 that the model training error is closely related to the initial model selection, mini-batch size B, the number of local iterations between consecutive coordination rounds E, and the total number of receivers participating in the model training K^L. More specifically, $\|\boldsymbol{\Omega}^0 - \boldsymbol{\Omega}^*\|^2$ term in (25) quantifies the error caused by the selection of the initial model. Since this term is multiplied with term $e^{\frac{-\tilde{\eta}_1\mu T}{4}}$, we can reduce the impact of incorrect selection of the initial model by increasing the

value of $\tilde{\eta}_1$ which can be achieved by choosing a smaller value λ. We can also observe that the model training error always increases with the values of B, E, and K^L. Increasing these parameters however will result in higher computation of complexity and longer coordination delay during each coordination round.

It is known that, in most existing FL-based solutions, the convergence rate is always adversely affected by the heterogeneity level of datasets at the model training participating receivers [40], [42]. Our result in Theorem 1 can also capture this issue. More specifically, the term Γ^L is a commonly used metric to measure the heterogeneity level, i.e., level of non-iid, among datasets at the receivers. We can observe that model training error increases with the value of Γ^L. We can however observe that the impact of the non-iid decreases when the number of coordination rounds T becomes large.

7.2 Model Transfer Error

In the model transfer step, receiver $k'' \in \mathcal{K}^N$ can directly obtain its model by aggregating the already trained models of others, e.g., receiver k' for $k' \in \mathcal{K}^L$. Therefore, the model transfer error is closely related to the data distribution difference between receivers in sets $\mathcal{K}^L$ and $\mathcal{K}^N$. In this paper, we use a commonly used metric, total variation distance, to quantify the data distribution difference which is defined as follows:

Definition 2. For the data distributions $\mathcal{P}_{k'}$ and $\mathcal{P}_{k''}$ of receivers k' and k'' over the dataset $\mathcal{D}$, the total variation distance between them is defined as $\|\mathcal{P}_{k'} - \mathcal{P}_{k''}\|_{TV} \coloneqq \sup_{\zeta\in\mathcal{D}}|\mathcal{P}_{k'}(\zeta) - \mathcal{P}_{k''}(\zeta)|$, where ζ is data uniformly sampled from dataset $\mathcal{D}$.

We can then prove the following result about the model transfer error.

Theorem 2. Suppose Assumptions 1-4 and the following assumptions hold:

Assumption 5: The local objective function is M-bounded: $F_k(\cdot, \zeta_k) \leq M$, for all $k \in \mathcal{K}$.

Then, we have

$$\mathbb{E}[F_{k''}(\boldsymbol{\omega}_{k''}) - F_{k''}(\boldsymbol{\omega}_{k''}^*)] \leq \frac{\epsilon + (\lambda + \Gamma^N)K^L}{K^L} \quad (26)$$
$$+ M\sum_{k'\in\mathcal{K}^L}\|\xi_{k'',k'}\mathcal{P}_{k''} - \frac{\mathcal{P}_{k'}}{K^L}\|_{TV},$$

where ϵ is the model training error derived in (25), $\phi_{k''}$ are defined in Sections 6.1 and 6, $\boldsymbol{\omega}_{k''} = \sum_{k'\in\mathcal{K}^L}\xi_{k'',k'}\boldsymbol{\omega}_{k'}$, $\xi_{k'',k'} = S(g_{\delta*}(\phi_{k''}), g_{\delta*}(\phi_{k'}))$, and $\Gamma^N = \frac{1}{K^L}\sum_{k'\in\mathcal{K}^L}F_{k'}(\boldsymbol{\omega}_{k'}^*) - F_{k''}(\boldsymbol{\omega}_{k''}^*)$.

Proof: See Appendix B. $\qquad\square$

Assumptions 5 and 6 in Theorem 2 are reasonable in many transfer learning application scenarios [49], [50]. This is because receivers without labeled data rely on the transferred model to recognize specific human gestures when observing new testing data points. In the ideal scenario in which the transferred model can perfectly recognize the label of any new testing data point, the local objective function $F_k(\boldsymbol{\Omega}_{k''}^*, \zeta_k)$ will be minimized and also the gradient value $\nabla F_k(\cdot, \zeta_k)$ will approach zero. Even in

most non-ideal scenarios, the local objective function and its gradient based on any new testing data point need to be assumed to be bounded to derive any valid theoretical bounds.

We can observe that the theoretical bound of model transfer error depends mainly on the TV distance between data distributions of receivers k' and k'' as shown in (27). More specifically, as the TV distance between $\mathcal{P}_{k'}$ and $\mathcal{P}_{k''}$ reduces, the term in (27) approaches zero. We can therefore apply solutions developed in Section 6.1 to learn the optimal model correlation coefficients $\xi_{k'',k'}$ to minimize (27). We can also observe that the model transfer error is also related to the model training error. However, as the number of model training receivers becomes large, the impact of the model training error on the transfer error decreases, as shown in (26).

8 PERFORMANCE EVALUATION

8.1 Experimental Setup

Dataset: To evaluate the performance of SANSee, we conduct extensive experiments based on a public available wireless sensing dataset, Widar [12], consisting of 6 types of human gestures (e.g., Push-Pull, Sweep, Clap, Slide, Draw-O, and Draw-zigzag) recorded at three different environments: classroom, hall, and office. In each environment, an off-the-shelf Wi-Fi transmitter with one activated antenna and 6 receivers, each has three activated antennas, have been deployed at different locations with the same relative distances in a 2m×2m sensing area. The transmitter is set to broadcast data packets at a rate of 1,000 packets per second at 5.825 GHz Wi-Fi band. The dataset consists of 12,000 labeled gesture data samples in total.

Model: For gesture recognition model construction, each gesture is assumed to last around 1.5 seconds and the CSI signals recorded by each receiver will be equally divided into a set of 1.5 second time segments. We adopt ResNet-8 model to extract the spatial and temporal features of the training data samples associated with different gestures, trained based on a cross-entropy loss function using the standard SGD algorithm. For the semantic-based transfer learning model, we design a 4-convolutional block-based CNN architecture to convert high-dimensional semantic features into low-dimensional semantic space in which each block consists of a 3×3 convolutional layer followed by a batch normalization and ReLU layer. To map semantic similarity to model correlation, the low-dimensional semantic representations are fed into a feature concatenation layer, followed by 2 convolutional blocks, a fully connected ReLU layer and a fully connected sigmod layer to output the model correlation coefficient.

Platform: We conduct our experiments on a workstation with an Intel(R) Core(TM) i9-13900K CPU@5.8GHz, 128.0GB RAM@4000.0MHz, 1 TB SSD, 4 TB HDD, and two NVIDIA GeForce RTX 4090 GPUs. The CSI data samples are processed using MatLab and gesture recognition models and pSAN-based transfer learning models are trained using Python 3.8, CUDA 12.2 and Pytorch 2.1.0 running on Ubuntu 22.04.

8.2 Physical-layer Semantics Estimation

SANSee is built based on the basic idea that the physical-layer semantics play a key role in determining the distributions of wireless sensing signals as well as the models to recognize different gestures. We therefore need to first evaluate the E- and G-semantics estimated by our proposed physical-layer semantics estimation algorithm under different settings.The main idea of human gesture recognition is to detect the impact of the Doppler shift caused by human body movements on the wireless signal, particularly its higher frequency content. In this case, the magnitude of the Doppler shift of wireless signals detected by the receiver mainly depends on the gesture-performing speed as well as the signal frequency for gesture detection. It is known that body movement speeds for most human gestures, including Sweep, Clap, and Slide considered in this paper, are between 0.25 m/sec and 4 m/sec [51], which correspond to the Doppler frequency shift between 8 Hz and 134 Hz at 5 GHz band [52]. We therefore set the threshold for separating the high-pass and low-pass filters to 2 Hz. In Fig. 4-6, we present the physical-layer semantic features, including amplitude, ToF, AoA, and DFS of both E- and G-semantics, estimated based on Algorithm 1 proposed in Section 5. We also show estimated results of the primary path, which responds to the reflected signal with the highest amplitude (navy blue points), with (red solid lines) and without (black dash lines) the Gaussian smoother (GS). These different features can be influenced by different semantic features of the environmental layout and gestures. For example, signal amplitudes are mainly affected by the transmit signal power as well as various power losses caused by environmental reflections, blockages, transmission distance between the transmitter and receivers. AoAs of the received signals are mainly affected by the relative orientations of the transmitter, receivers, and the gesture performing human user. In Fig. 4-6, we can observe that the impact of these semantic features can be perfectly captured by the stationary and dynamic path components estimated by our proposed algorithm. For example, in Fig. 4-5, due to the differences in movement patterns, we can observe that all estimated G-semantics parameters of gestures "Push & Pull" and "Sweep" are significantly dissimilar to each other. E-semantics of "Push & Pull" and "Sweep" gestures look very similar as they are recorded in the same environment office and also the three main components observed in the amplitudes of E-semantics correspond to the signals received from the direct path and two paths reflected from the walls. Moreover, in Fig. 5-6, we can observe that the fluctuation patterns of the G-semantic parameters of gesture "Sweep" look very similar to each other even they are performed in the different environments, but the G-semantics are different due to the change in the physical environments. This also suggests that neither E-semantic nor G-semantic alone will not be able to capture the full picture of the impact of human gestures on wireless signals. Generally speaking, taking into consideration more physical-layer semantic features will result in a higher gesture recognition accuracy. It may however result in a higher computational complexity as will be discussed next.
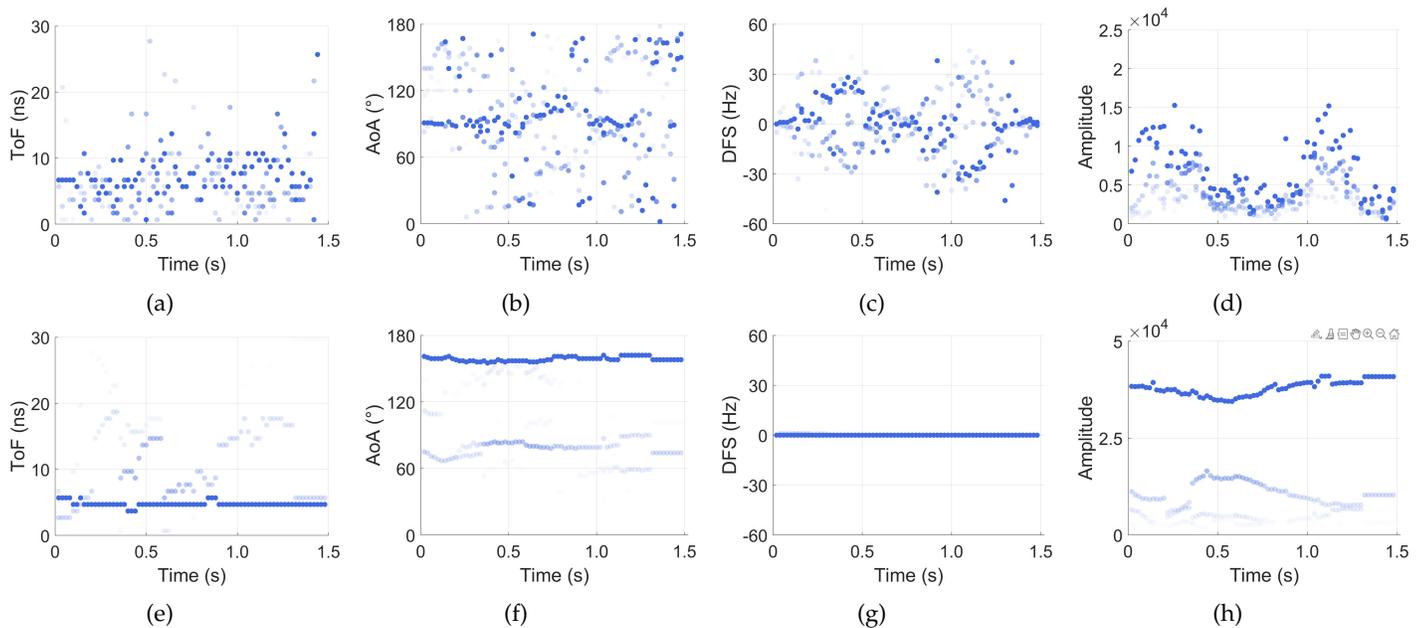
Fig. 4: G-semantic features (b)-(e) and E-semantic features (g)-(j) of gesture "Push & Pull" (a) performed in the office environment (f), based on $L = 5$ estimated paths.
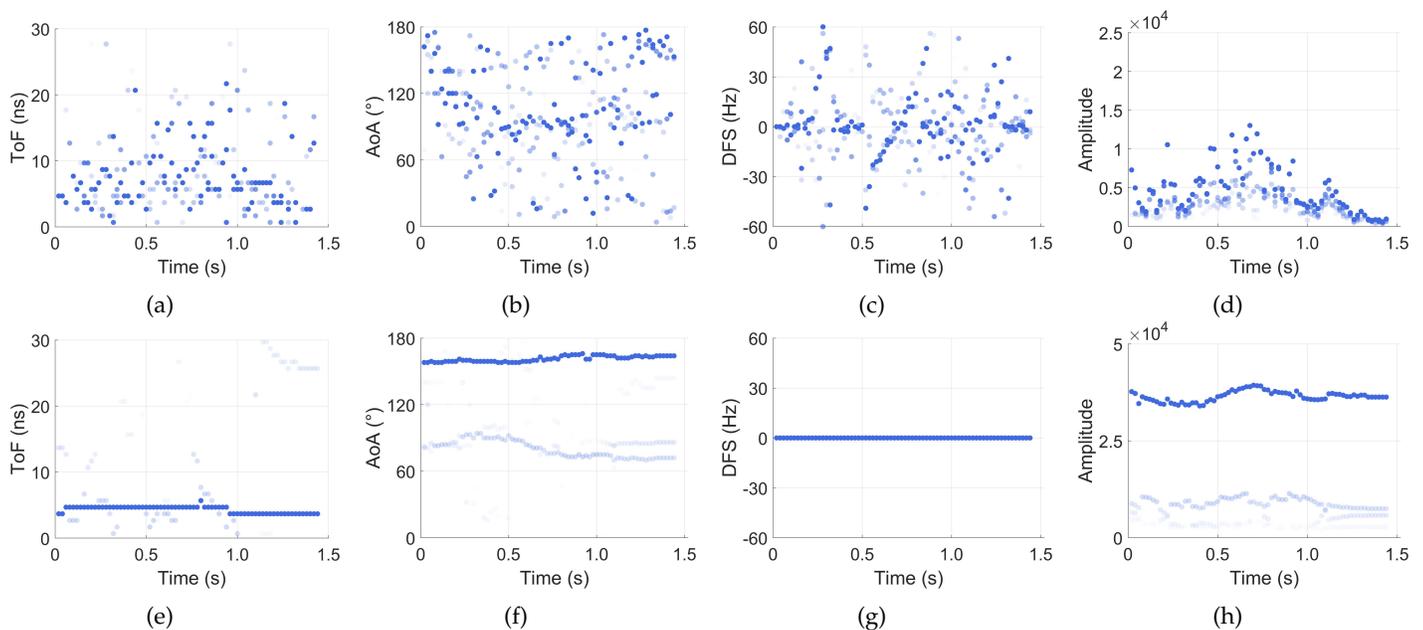

Fig. 5: G-semantic features (b)-(e) and E-semantic features (g)-(j) of gesture "Sweep" (a) performed in the office environment (f), based on $L = 5$ estimated paths.

Based on the above observations, we then evaluate the impact of different physical-layer semantic features on the recorded CSI signals at the receivers. In Fig. 7, we present the t-SNE-based visualizations of the statistical features of CSI signals of the same gesture recorded at different locations in different environments. We observe that, even the relative locations and orientations of the transmitter, receivers, and the human user remain the same at different environments, the recorded CSI signals may vary significantly. This further justifies our observations that the traditional centralized modeling approaches, in which

wireless sensing data samples recorded at different locations are combined at a centralized server to train a single global model for recognizing gestures performed at different locations, cannot provide accurate and consistent wireless sensing results at different receivers, especially in complex environments.

8.3 Model Training

To evaluate the extra computational complexity introduced by considering more semantic features in the model training, we use the time consumption of training a given
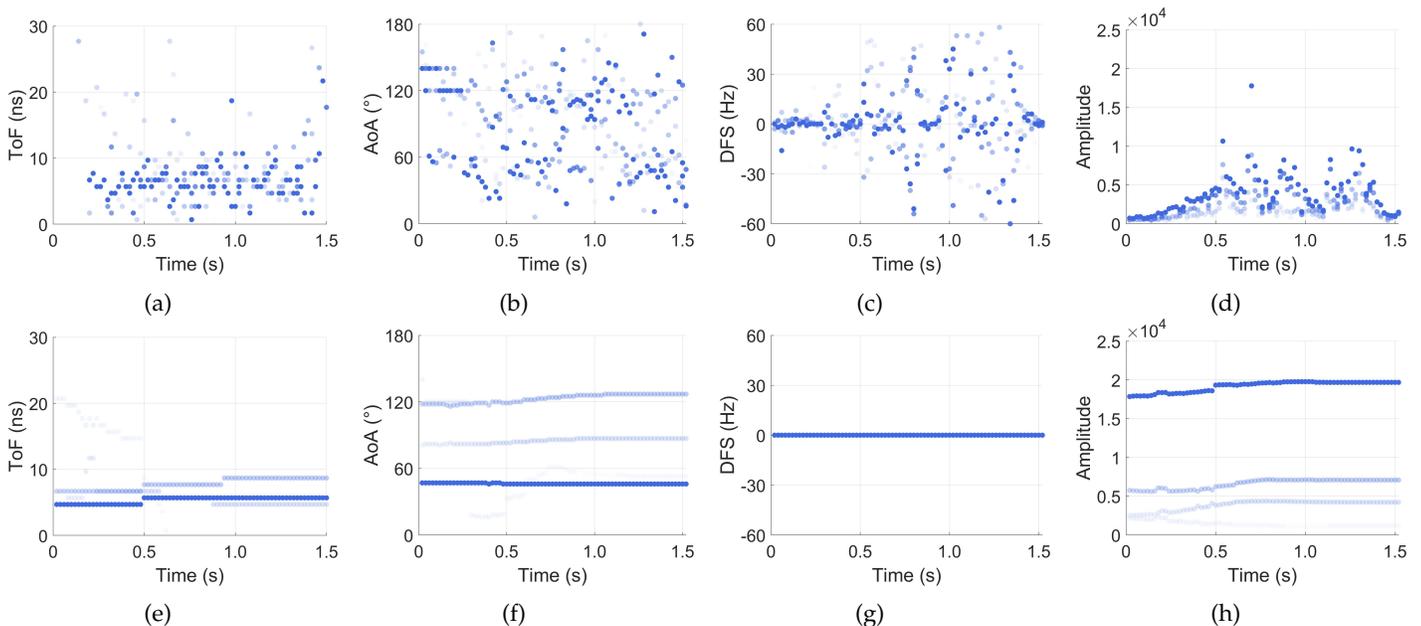
Fig. 6: G-semantic features (b)-(e) and E-semantic features (g)-(j) of gesture "Sweep" (a) performed in the classroom environment (f), based on $L = 5$ estimated paths.
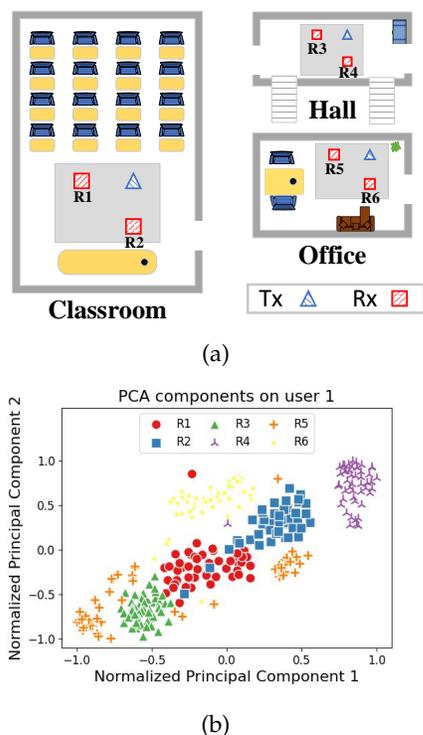

(a)

(b)

Fig. 7: (a) Locations of 6 receivers (labeled as R1-R6) deployed in 3 different environments, and (b) t-SNE-based visualization of statistical diversity of the CSIs of the same gesture recorded by different receivers.

model with a fixed number of iterations as the main metric to evaluate the complexity of model training, and compare the required model training time and the resulting model accuracy when different combinations of semantic features are fed into the model during training in Table I. We can observe that the model training time almost doubles when

a new semantic feature, E- or G-semantics, is added in the model training. Despite the increase in model training complexity, the accuracy of gesture recognition improves significantly, e.g., the gesture recognition accuracy improves over 50%, increasing from 61.67% with only amplitudes being considered to 96.34% with all the features of both G- and E-semantics being included in the model training. We also evaluate the impact of estimating different numbers of path components in Algorithm 1 on the model complexity and accuracy in Table I. We can observe that, when the number of estimated path components increases from $L = 5$ to $L = 20$, the overall time consumption increases only at around 8% and the resulting model accuracy improves around 18%.

Let us now evaluate the model training performance of SANSee for the receivers with labelled data. We compare the model accuracy of all 18 receivers at three different environments achieved by SANSee to the state-of-the-art algorithms in Fig. 8. More specifically, in addition to comparing SANSee with the local training (*Local*) in which each receiver trains a local model based only on its local dataset and *FedAvg* [32] in which all receivers train a single global model by periodically aggregating their local model parameters, we also consider three state-of-art personalized federated learning algorithms: *pFedMe* [40], *FedAMP* [47], and *Ditto* [41]. Moreover, Fig. 8 includes the average accuracy calculated based on models trained at 10 experiments. We also highlight the highest and lowest bounds on accuracy for models trained at different receivers. In Fig. 8, we can observe that SANSee outperforms all these existing personalized model training algorithms and can achieve model accuracy improvements between 9.44 % and 27.64 % on average. Furthermore, the model performance at different receivers is more consistent in SANSee compared to other algorithms. More specifically, in local training, FedAvg, pFedMe, FedAMP, and Ditto algorithms, the gap between the highest and lowest model

	Amplitude Only (L=10)	DFS Only (L=10)	G-semantics Only (L=10)	E-semantics Only (L=10)	Both E- and G-Semantics (L=5)	Both E- and G-Semantics (L=10)	Both E- and G-Semantics (L=20)
Dimensional Size of Semantics	$1 \times 100 \times 300$	$1 \times 120 \times 300$	$4 \times 120 \times 300$	$4 \times 120 \times 300$	$8 \times 120 \times 300$	$8 \times 120 \times 300$	$8 \times 120 \times 300$
Model Training Time	1.78 h	2.23 h	6.04 h	6.04 h	13.37 h	13.90 h	14.44 h
Model Accuracy	61.67%	73.87%	89.98%	16.67%	81.11%	92.90%	96.34%

TABLE 1: Comparison of model training time and accuracy when considering different combinations of semantic features and the numbers of estimated path components.

accuracy when implementing the trained models at different receivers are 18.75%, 11.63%, 14.02%, 12.20%, 22.76%, respectively, all of which is larger than the 9.16% gap achieved by our proposed SANSee.

To verify the theoretical results derived in Section 7.1, we evaluate the convergence performance of the model training process at receivers with labelled data under different number of coordination rounds and combinations of key model parameters including λ, σ_R, K^L, B and E in Fig. 9 and 10.

Recall that λ is the collaboration parameter that controls the weights of the attention-inducing regulation function in the local objective function of each receiver. Increasing λ accelerates collaboration between receivers with highly correlated models. We can observe in Fig. 8(a) that when the value of λ increases from zero to one, the model convergence speed also increases. However, when λ continues to increase from one to 10, the model accuracy will be degraded. This is because λ can only control the weight of the regularization term in the local objective function, and when this weight becomes too high, the regularization term will overwhelm the overall local objective function, resulting in high distortion on the original local objective as well as resulting models. Therefore, there is an optimal λ for the target problem, which can not only accelerate the model convergence and avoid overfitting, but also prevent the regularization term from overwhelming the effect of the cross-entropy loss term because the large penalty of increasing the weight modulus from 0 distorts the shape of the loss surface.

Similarly, in Fig. 9(b), we can observe that another key parameter σ_R in the negative exponential regularization function to control the weights of aggregation of correlated model also needs to be carefully chosen to improve the model accuracy level with maximized convergence speed, e.g., as observed in Fig. 9(b), the highest convergence performance is achieved when $\sigma_R = 1$. In the rest of this section, we set both values of λ and σ_R into 1. In Fig. 9(c), we present the convergence rate under different numbers of receivers participating in the model training. It is known that for most traditional federated learning solutions, if datasets at different receivers are non-iid, allowing more receivers to participate in the model training generally results in reduced convergence rates. We can observe in Fig. 9(c), however, that the convergence performances of SANSee do not change much even when the number of receivers participating in the model training increases from 2 to 18.

In Fig. 10, we compare the average model accuracy and the loss values under different coordination rounds and combinations of mini-batch sizes B and local iteration (epoch) numbers between consecutive coordination rounds E. We can observe in Theorem 1 that the convergence rate

is in the order of $\mathcal{O}(\frac{1}{E})$ when all the other parameters are fixed, which is aligned with Fig. 10(a) and (b), in which we can observe that, as E increases from 1 to 10, the number of global coordination rounds also increases. Similarly, in Fig. 10(c) and (d), we fix $E = 5$ and compare the convergence performance of SANSee under different B. We can observe that increasing B results in almost linear reduction of the required number of coordination rounds T to convergence.

In Fig. 11(a), we compare the performance of wireless sensing when adopting different models in SANSee, including a 2-layer CNN, a lightweight CNN called Mobilenet-v2 [53], a hybrid neural network (CNN+GRU) consisting of a CNN for spatial feature extraction and a GRU for temporal modeling [12], as well as ResNet-8, ResNet-18, and ResNet-50. We also compare the computational complexity (in FLOPs) and sizes of parameters of these models in Fig. 11(b). We observe that, in terms of average accuracy, ResNet-18 outperforms the other models, achieving 2.31% to 7.45% improvements on average. When considering the variance of wireless sensing, however, adopting more complex models (with a higher number of parameters) can always reduce the variance. This is due to the fact that for a given number of training samples, models with small or large numbers of parameters tend to cause underfitting or overfitting issues, resulting in higher bias with lower variance or lower bias with higher variance in performance. Furthermore, we can observe that lightweight models such as Mobilenet-v2 and ResNet-8 can still achieve relatively good wireless sensing accuracy. Furthermore, choosing complex models such as ResNet-50, i.e., models with large numbers of parameters, may not always result in improved performance. This is because large models may result in overfitting.

In Fig. 12, we compare the average model accuracy of SANSee under different numbers of gesture classes (Fig. 12(a)) and different training dataset sizes per gesture (with six gestures in total) (Fig. 12(b)) at each receiver. From Fig. 12(a), we observe that when the number of gesture classes increases from 2 to 9, the average model accuracy decreases from 99.809% to 84.915%. This is because, as the number of gesture classes increases, the output dimension of the model also increases, resulting in underfitting issues for each class of gestures. In Fig. 12(b), we can observe that when the training dataset size per gesture at each receiver increases, the increasing speed of the average model accuracy decreases. For example, as the number of samples per gesture at each receiver increases from 1 to 25, the model accuracy increases from 49.05% to 87.28%, resulting in 38.23% improvement. However, if the training sample size continues to increase from 75 to 100, the model accuracy improves by only 0.417%.
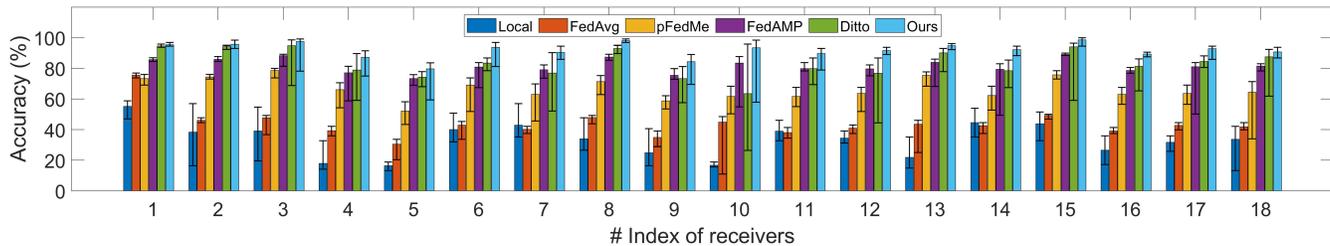
Fig. 8: Wireless sensing accuracy at 18 receivers achieved by models trained by different algorithms, including local training, FedAvg, pFedMe, FedAMP, Ditto, and the proposed SANSee.
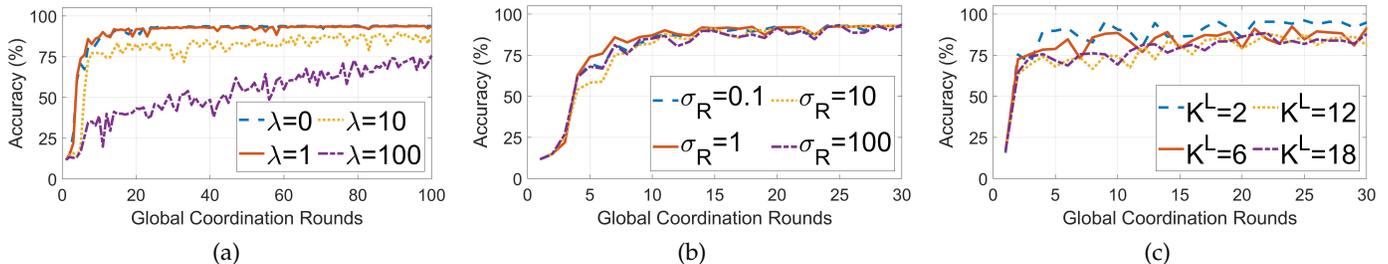

(a)

(b)

(c)

Fig. 9: Comparison of convergence rates under different model training parameters, including (a) λ ($\sigma_R = 1, K^L = 18$); (b) σ_R ($\lambda = 1, K^L = 18$); and (c) K^L ($\lambda = 1, \sigma_R = 1$).
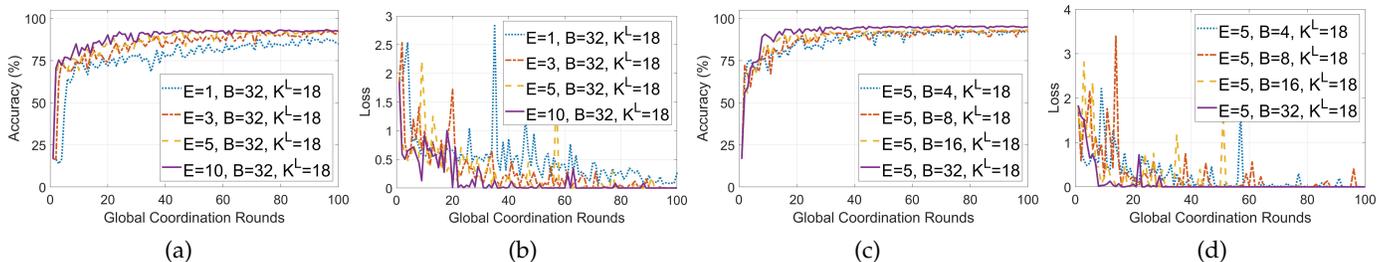

(a)

(b)

(c)

(d)

Fig. 10: Comparison of convergence rates of model training under different batch-sizes and local epoch numbers.
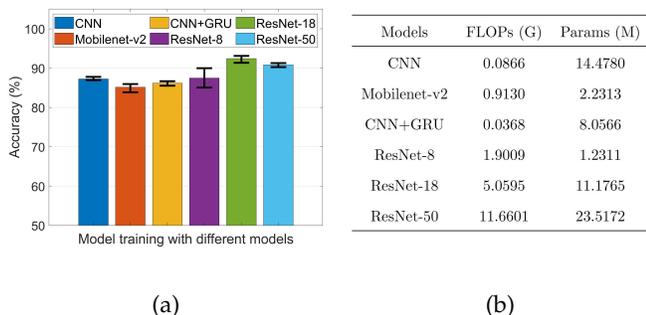

(a)

(b)

Fig. 11: (a): Model accuracies achieved by different models, each with computational complexity (in FLOPs) and size of parameters listed in Figure (b), including a 2-layer CNN network (CNN), a lightweight CNN network (Mobilenet-v2) [53], a hybrid neural network composed of CNN and GRU (CNN+GRU) [12], ResNet-8, ResNet-18, and ResNet-50.
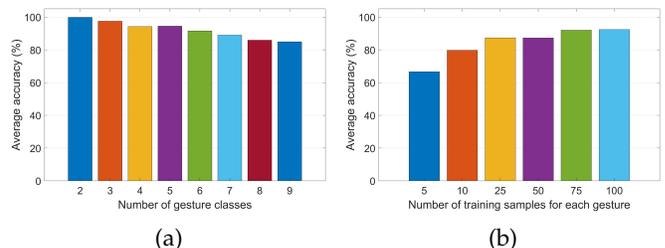

(a)

(b)

Fig. 12: Average model accuracy with (a) different numbers of gesture classes, and (b) different training dataset sizes per gesture (with six gestures in total) at each receiver.

8.4 Model Transfer

To evaluate the performance of SANSee for transferring the already trained models, i.e., base models, to receivers without labeled dataset, we consider two model transfer scenarios: *in-environment model transfer* in which all the receivers with and without labelled data are in the same environment and *cross-environment model transfer* in which models trained by receivers in one environment are transferred to receivers located in a new environment.

We evaluate the in-environment model transfer performance in Fig. 14, in which we compare the gesture recognition accuracy of models obtained by a receiver without labeled data when its model are transferred based on different numbers, 1-4, of available models constructed by receivers with labeled data. We can observe that, as the

number of available models increases, the accuracy of the transferred model also improves. The increasing speed of the model accuracy however decreases as the number of available models becomes large. This means that SANSee is able to transfer a relatively small number of trained models, e.g., trained by two to three receivers, to any number of location-specific models with sufficiently "good" accuracy, e.g., above 70% gesture recognition accuracy.
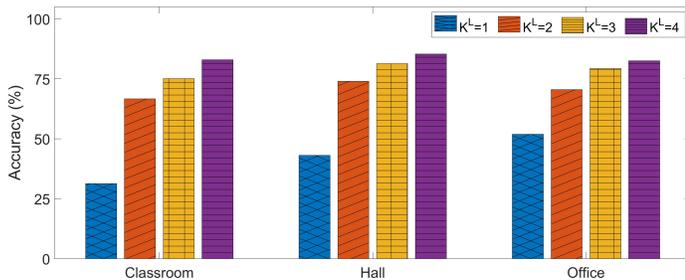

Fig. 13: Comparison of in-environment model transfer performance at three environment under different numbers of base models K^L.

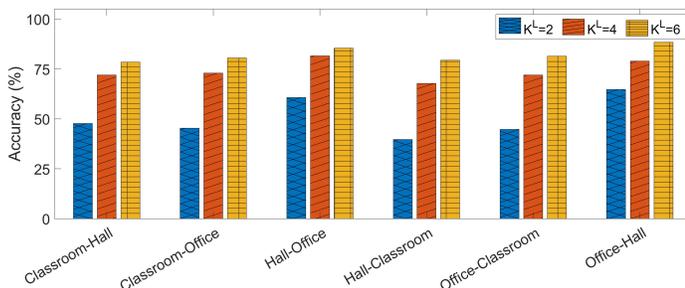Fig. 14: Comparison of cross-environment model transfer performance under different numbers of base models K^L.

We evaluate the cross-environment model transfer performance in Fig. 13 where models trained by receivers in one environment are transferred to the receivers in another environment. We can observe that, generally speaking, the accuracy of the cross-environment model transfer is slightly worse than that of the in-environment model transfer under the same number of available models. The performance of the transferred model is again affected by the number of models that have already been trained. For example, in hall-classroom cross-environment model transfer scenario, as the number of models increases from 2 to 6, the accuracy of the transferred model improves almost 50% from accuracies 41% to 83%. The increasing speed of the model performance again approaches a stationary level when the number of available models increases. In other words, SANSee is a useful solution for achieving sustainable network AI in a large networking system, in which an almost infinite number of novel models tailored for a wide range of applications and scenarios can be transferred based on a very limited number of base models using their semantic correlations.

9 CONCLUSION

This paper proposed a semantic-aware networking-based framework for distributed wireless sensing, called SANSee, that allowed models constructed in a limited number of locations to be directly transferred to other locations without any training efforts. In particular, a physical-layer semantic-aware network, called pSAN, has been developed to characterize the similarity between physical-layer semantic features and the correlations of wireless sensing data distributions across different locations. We have then proposed a pSAN-based zero-shot transfer learning solution for receivers without labeled data to construct its location-specific model based on the correlated model trained by receivers with labeled data. Finally, extensive experiments have been conducted based on real-world datasets to evaluate the performance of SANSee, and numerical results showed the accuracy of transferred models obtained by SANSee matched that of the models trained by the locally labeled data based on supervised learning approaches.

REFERENCES

[1] H. Zhu, Y. Xiao, Y. Li, G. Shi, and W. Saad, "Physical-layer semantic-aware network for zero-shot wireless sensing," in *IEEE ICNP Workshop*, pp. 1–6, Reykjavik, Iceland, Oct. 2023.

[2] Y. Xiao, G. Shi, Y. Li, W. Saad, and H. V. Poor, "Toward self-learning edge intelligence in 6G," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 34–40, Dec. 2020.

[3] Y. Xiao and M. Krunz, "Distributed optimization for energy-efficient fog computing in the tactile Internet," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2390–2400, Nov. 2018.

[4] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Commun. Surv. Tut.*, vol. 22, no. 3, pp. 1629–1645, Aug. 2019.

[5] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *J. Mach. Learn. Res.*, vol. 9, no. 11, Nov. 2008.

[6] Y. Ma, G. Zhou, and S. Wang, "Wifi sensing with channel state information: A survey," *ACM Comput. Surv.*, vol. 52, no. 3, pp. 1–36, Jun. 2019.

[7] S. Tan, Y. Ren, J. Yang, and Y. Chen, "Commodity wifi sensing in ten years: Status, challenges, and opportunities," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17832–17843, Apr. 2022.

[8] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using wifi channel state information," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 98–104, Oct. 2017.

[9] Z. Wang, K. Jiang, Y. Hou, W. Dou, C. Zhang, Z. Huang, and Y. Guo, "A survey on human behavior recognition using channel state information," *IEEE Access*, vol. 7, pp. 155986–156024, Oct. 2019.

[10] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "Spotfi: Decimeter level localization using wifi," in *ACM SIGCOMM*, p. 269–282, United Kingdom, London, Aug. 2015.

[11] X. Li, D. Zhang, Q. Lv, J. Xiong, S. Li, Y. Zhang, and H. Mei, "Indotrack: Device-free indoor human tracking with commodity wi-fi," *ACM IMWUT*, vol. 1, no. 3, pp. 1–22, Sept. 2017.

[12] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wi-fi," in *ACM MobiSys*, pp. 313–325, Seoul, Korea, Jun. 2019.

[13] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Commun. Mag.*, vol. 59, no. 8, pp. 44–50, Aug. 2021.

[14] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Nov. 2023.

[15] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2434–2444, Jun. 2021.

[16] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2584–2597, Jul. 2022.

[17] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, Apr. 2021.

[18] J. Chai, Y. Xiao, G. Shi, and W. Saad, "Rate-distortion-perception theory for semantic communication," in *IEEE ICNP*, pp. 1–6, Reykjavik, Iceland, Oct. 2023.

[19] Y. Xiao, X. Zhang, Y. Li, G. Shi, and T. Başar, "Rate-distortion theory for strategic semantic communication," in *IEEE ITW*, pp. 279–284, Mumbai, India, Dec. 2022.

[20] Y. Xiao, Z. Sun, G. Shi, and D. Niyato, "Imitation learning-based implicit semantic-aware communication networks: Multi-layer representation and collaborative reasoning," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 3, pp. 639–658, Dec. 2022.

[21] Y. Xiao, Y. Liao, Y. Li, G. Shi, H. V. Poor, W. Saad, M. Debbah, and M. Bennis, "Reasoning over the air: A reasoning-based implicit semantic-aware communication framework," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, Apr. 2024.

[22] Y. Yang, F. Gao, X. Tao, G. Liu, and C. Pan, "Environment semantics aided wireless communications: A case study of mmwave beam prediction and blockage prediction," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 2025–2040, Jul. 2023.

[23] J. Feng, S. Li, X. Li, F. Wu, Q. Tian, M.-H. Yang, and H. Ling, "Taplab: A fast framework for semantic video segmentation tapping into compressed-domain knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1591–1603, Sept. 2020.

[24] Z. Liu, C. Wen, Z. Su, S. Liu, J. Sun, W. Kong, and Z. Yang, "Emotion-semantic-aware dual contrastive learning for epistemic emotion identification of learner-generated reviews in moocs," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, Jul. 2023.

[25] H. Deng, Z. Yang, T. Hao, Q. Li, and W. Liu, "Multimodal affective computing with dense fusion transformer for inter-and intra-modality interactions," *IEEE Trans. Multimedia*, vol. 25, pp. 6575–6587, Sept. 2022.

[26] J. Yang, X. Chen, H. Zou, C. X. Lu, D. Wang, S. Sun, and L. Xie, "Sensefi: A library and benchmark on deep-learning-empowered wifi human sensing," *Patterns*, vol. 4, no. 3, p. 100703, Mar. 2023.

[27] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, *et al.*, "Towards environment independent device free human activity recognition," in *ACM MobiCom*, pp. 289–304, New Delhi, India, Oct. 2018.

[28] H. Zou, J. Yang, Y. Zhou, L. Xie, and C. J. Spanos, "Robust wifi-enabled device-free gesture recognition via unsupervised adversarial domain adaptation," in *ICCCN*, pp. 1–8, IEEE, Hangzhou, China, Oct. 2018.

[29] D. Wang, J. Yang, W. Cui, L. Xie, and S. Sun, "Airfi: empowering wifi-based passive human gesture recognition to unseen environment via domain generalization," *IEEE Trans. Mob. Comput.*, vol. 23, no. 2, pp. 1156–1168, Feb. 2024.

[30] C. Li, Z. Cao, and Y. Liu, "Deep ai enabled ubiquitous wireless sensing: A survey," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–35, Mar. 2021.

[31] J. Zhang, Z. Tang, M. Li, D. Fang, P. Nurmi, and Z. Wang, "Crosssense: Towards cross-site and large-scale wifi sensing," in *ACM MobiCom*, pp. 305–320, New Delhi, India, Oct. 2018.

[32] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *PMLR Artificial Intelligence and Statistics*, pp. 1273–1282, Seoul, Korea, Nov. 2017.

[33] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Commun. Surv. Tutor.*, vol. 23, no. 3, pp. 1622–1658, Apr. 2021.

[34] S. M. Hernandez and E. Bulut, "Wifederated: Scalable wifi sensing using edge-based federated learning," *IEEE Internet Things J.*, vol. 9, no. 14, pp. 12628–12640, Jul. 2022.

[35] K. Zhang, X. Liu, X. Xie, J. Zhang, B. Niu, and K. Li, "A cross-domain federated learning framework for wireless human sensing," *IEEE Netw.*, vol. 36, no. 5, pp. 122–128, Nov. 2022.

[36] Y. Liu, H. Li, J. Xiao, and H. Jin, "Floc: Fingerprint-based indoor localization system under a federated learning updating framework," in *MSN*, pp. 113–118, IEEE, Shenzhen, China, Apr. 2019.

[37] N. Nagia, M. T. Rahman, and S. Valaee, "Federated learning for wifi fingerprinting," in *IEEE ICC*, pp. 4968–4973, IEEE, Seoul, Korea, Aug. 2022.

[38] K. Qian, C. Wu, Y. Zhang, G. Zhang, Z. Yang, and Y. Liu, "Widar2. 0: Passive human tracking with a single wi-fi link," in *ACM MobiSys*, pp. 350–361, Munich, Germany, Jun. 2018.

[39] B. Fleury, M. Tschudin, R. Heddergott, D. Dahlhaus, and K. Ingeman Pedersen, "Channel parameter estimation in mobile radio environments using the sage algorithm," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 3, pp. 434–450, Mar. 1999.

[40] C. T Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with moreau envelopes," in *NeurIPS*, vol. 33, pp. 21394–21405, Virtual. Dec. 2020.

[41] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *PMLR ICML*, vol. 139, pp. 6357–6368, Virtual, Jul. 2021.

[42] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *MLSys*, pp. 429–450, Austin, USA, Mar. 2020.

[43] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *PMLR ICML*, vol. 119, pp. 5132–5143, Virtual, Jul. 2020.

[44] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *CVPR*, pp. 10713–10722, Virtual, Jun. 2021.

[45] C. T. Dinh, T. T. Vu, N. H. Tran, M. N. Dao, and H. Zhang, "A new look and convergence rate of federated multitask learning with laplacian regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 1–11, Dec. 2022.

[46] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 9587–9603, Dec. 2023.

[47] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei, and Y. Zhang, "Personalized cross-silo federated learning on non-iid data," in *AAAI*, vol. 35, pp. 7865–7873, Virtual, May. 2021.

[48] A. Gouissem, Z. Chkirbene, and R. Hamila, "A comprehensive survey on energy efficiency in federated learning: Strategies and challenges," in *ENERGYCON*, pp. 1–6, vibrant city, Doha, Qatar, Apr. 2024.

[49] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning: A meta-learning approach," *arXiv preprint arXiv:2002.07948*, 2020.

[50] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks," *NeurIPS*, vol. 34, pp. 5469–5480, Virtual, Dec. 2021.

[51] S. Gupta, D. Morris, S. Patel, and D. Tan, "Soundwave: using the doppler effect to sense gestures," in *ACM CHI*, p. 1911–1914, Austin, Texas, USA, May. 2012.

[52] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *ACM MobiCom*, pp. 27–38, Miami, Florida, USA, Sep. 2013.

[53] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, pp. 4510–4520, Salt Lake City, Utah, USA, Jun. 2018.
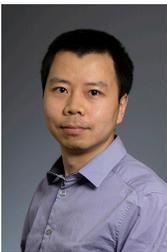
Huixiang Zhu received his B.S. degree in Wuhan University of Technology, Wuhan, China in 2019. He is currently pursuing his PhD degree in the School of Electronic Information and Communications at the Huazhong University of Science and Technology, Wuhan, China. His research interests include semantic-aware communications and network AI.

Guangming Shi (Fellow, IEEE) received the M.S. degree in computer control and the Ph.D. degree in electronic information technology from Xidian University, Xi'an, China, in 1988, and 2002, respectively. He was the vice president of Xidian University from 2018 to 2022. Currently, he is the Vice Dean of Peng Cheng Laboratory and a Professor with the School of Artificial Intelligence, Xidian University. He is an IEEE Fellow, the chair of IEEE CASS Xi'an Chapter, senior member of ACM and CCF, Fellow of Chinese Institute of Electronics, and Fellow of IET. He was awarded Cheung Kong scholar Chair Professor by the ministry of education in 2012. He won the second prize of the National Natural Science Award in 2017. His research interests include Artificial Intelligence, Semantic Communications, and Human-Computer Interaction.

Yong Xiao (Senior Member, IEEE) received his B.S. degree in electrical engineering from China University of Geosciences, Wuhan, China in 2002, M.Sc. degree in telecommunication from Hong Kong University of Science and Technology in 2006, and his Ph. D degree in electrical and electronic engineering from Nanyang Technological University, Singapore in 2012. He is now a professor in the School of Electronic Information and Communications at the Huazhong University of Science and Technology (HUST), Wuhan, China. He is also with Peng Cheng Laboratory, Shenzhen, China and Pazhou Laboratory (Huangpu), Guangzhou, China. He is the associate group leader of the network intelligence group of IMT-2030 (6G promoting group) and the vice director of 5G Verticals Innovation Laboratory at HUST. Before he joins HUST, he was a research assistant professor in the Department of Electrical and Computer Engineering at the University of Arizona where he was also the center manager of the Broadband Wireless Access and Applications Center (BWAC), an NSF Industry/University Cooperative Research Center (I/UCRC) led by the University of Arizona. His research interests include machine learning, game theory, distributed optimization, and their applications in semantic communications and semantic-aware networks, cloud/fog/mobile edge computing, green communication systems, wireless communication networks, and Internet-of-Things (IoT).

Marwan Krunz (Fellow, IEEE) is a Regents Professor at the University of Arizona. He holds the Kenneth VonBehren Endowed Professorship in ECE and is also a professor of computer science. He directs the Broadband Wireless Access and Applications Center (BWAC), a multi-university NSF/industry center that focuses on next-generation wireless technologies. He also holds a courtesy appointment as a professor at University Technology Sydney. Previously, he served as the site director for Connection One, an NSF/industry-funded center of five universities and 20+ industry affiliates. Dr. Krunz's research is in the fields of wireless communications, networking, and security, with recent focus on applying AI and machine learning techniques for protocol adaptation, resource management, and signal intelligence. He has published more than 320 journal articles and peer-reviewed conference papers, and is a named inventor on 12 patents. His latest h-index is 60. He is an IEEE Fellow, an Arizona Engineering Faculty Fellow, and an IEEE Communications Society Distinguished Lecturer (2013-2015). He received the NSF CAREER award. He served as the Editor-in-Chief for the IEEE Transactions on Mobile Computing. He also served as editor for numerous IEEE journals. He was the TPC chair for INFOCOM'04, SECON'05, WoWMoM'06, and Hot Interconnects 9. He was the general vice-chair for WiOpt 2016 and general co-chair for WiSec'12. Dr. Krunz served as chief scientist/technologist for two startup companies that focus on 5G and beyond wireless systems.

Yingyu Li (Member, IEEE) received the B.Eng. degree in electronic information engineering and the Ph.D. degree in circuits and systems from the Xidian University, Xi'an, China, in 2012 and 2018, respectively. From 2014 to 2016, she was a Research Scholar with the Department of Electronic Computer Engineering at the University of Houston, TX, USA. She was a postdoctoral researcher in the School of Electronic Information and Communications at Huazhong University of Science and Technology from 2018 to 2021. She is now an associate professor at the School of Mechanical Engineering and Electronic Information, China University of Geosciences (Wuhan). Her research interests include semantic communications, edge intelligence, green communication networks, and IoT.

APPENDIX A
PROOF OF THEOREM 1

In this section, we present detailed derivation of the (local) model training error $\mathbb{E}[\mathcal{J}(\boldsymbol{\Omega}) - \mathcal{J}(\boldsymbol{\Omega}^*)]$ in Theorem 1, where $\mathcal{J}(\boldsymbol{\Omega})$ is defined previously in (18), can converge to near the minimum as the number of training rounds T increases. Before we present the detailed proofs, let us first introduce the following lemmas which will be useful for our proofs of Theorem 1.

A.1 Key Lemmas of Theorem 1

Recall the definition of $\mathcal{J}(\boldsymbol{\Omega}) = \mathcal{F}(\boldsymbol{\Omega}) + \lambda\mathcal{R}(\boldsymbol{\Omega})$ in (18), where $\mathcal{R}(\boldsymbol{\Omega}) = \sum_{k'\in\mathcal{K}^L}\sum_{k'\neq j'} R\left(\|\boldsymbol{\omega}_{k'} - \boldsymbol{\omega}_{j'}\|^2\right)$ and $\mathcal{F}(\boldsymbol{\Omega}) = \sum_{k'\in\mathcal{K}^L} F_{k'}(\boldsymbol{\omega}_{k'})$, respectively.

Lemma 1. Suppose that Assumptions 1-3 in Theorem 1 hold. If $L_J := L + \lambda\kappa_R$, then for any $\boldsymbol{\Omega}, \boldsymbol{\Omega}' \in \mathbb{R}^{dK}$, we can derive the following results:
(a) $\|\nabla\mathcal{F}(\boldsymbol{\Omega}) - \nabla\mathcal{F}(\boldsymbol{\Omega}')\| \leq L\|\boldsymbol{\Omega} - \boldsymbol{\Omega}'\|$;
(b) $\|\nabla\mathcal{J}(\boldsymbol{\Omega}) - \nabla\mathcal{J}(\boldsymbol{\Omega}')\| \leq L_J\|\boldsymbol{\Omega} - \boldsymbol{\Omega}'\|$;
(c) $\|\nabla\mathcal{F}(\boldsymbol{\Omega})\|^2 \leq 2L^2\|\boldsymbol{\Omega} - \boldsymbol{\Omega}'\|^2 + 2\|\nabla\mathcal{F}(\boldsymbol{\Omega}')\|^2$;
(d) $\|\nabla\mathcal{J}(\boldsymbol{\Omega})\|^2 \leq 2L_J^2\|\boldsymbol{\Omega} - \boldsymbol{\Omega}'\|^2 + 2\|\nabla\mathcal{J}(\boldsymbol{\Omega}')\|^2$;
(e) $\mathcal{J}(\boldsymbol{\Omega}) - \mathcal{J}(\boldsymbol{\Omega}') \leq \langle\nabla\mathcal{J}(\boldsymbol{\Omega}'), \boldsymbol{\Omega} - \boldsymbol{\Omega}'\rangle + \frac{L_J}{2}\|\boldsymbol{\Omega} - \boldsymbol{\Omega}'\|^2$.
(f) $\mathbb{E}\|\nabla\widetilde{\mathcal{J}}(\boldsymbol{\Omega}, \zeta) - \nabla\mathcal{J}(\boldsymbol{\Omega})\|^2 \leq \sigma_F^2$

Lemma 2. Suppose that Assumption 1-2 hold and $\lambda > 2L$. Then there exists a positive value Γ^L such that, for any $\boldsymbol{\Omega}, \boldsymbol{\Omega} \in \mathbb{R}^{dK}$, we have

$$\|\nabla\mathcal{F}(\boldsymbol{\Omega}^*)\|^2 \leq \Gamma^L. \tag{27}$$

Γ^L is measured only at unique solution $\boldsymbol{\Omega}^*$, and thus Γ^L is finite. The bound is tight in the sense that $\Gamma^L = \lambda^2\|\nabla\mathcal{R}(\boldsymbol{\Omega}^*)\|^2 = 0$ for the i.i.d. cases, where $\boldsymbol{\omega}_1 = \cdots = \boldsymbol{\omega}_{K^L}$. In the considered wireless scenarios, there is always $\Gamma^L > 0$ due to the heterogeneity of wireless data samples.

Lemma 3. Suppose that Assumptions 4 holds. The gradient of server update in [21] at the tth coordination round, denoted as g^t, can be be represented as follows:

$$g^t = \sum_{e=0}^{E-1} \nabla\widetilde{\mathcal{J}}(\boldsymbol{\Omega}^{t,e}) + \lambda\kappa_R\mathcal{L}(\boldsymbol{\Omega}^{t,e} - \boldsymbol{\Omega}^t)$$
$$+ \frac{\eta\lambda\kappa_R\mathcal{L}}{E}\sum_{e=0}^{E-1} \nabla\widetilde{F}(\boldsymbol{\Omega}^{t,e}). \tag{28}$$

Proof: In the tth coordination round, the eth local update of the model vector $\boldsymbol{\Omega}$ in (20) can be represented as follows:

$$\boldsymbol{\Omega}^{t,e+1} = \boldsymbol{\Omega}^{t,e} - \eta\nabla\widetilde{F}(\boldsymbol{\Omega}^{t,e}) \tag{29}$$

implies that after E local update steps, we have

$$\eta\sum_{e=0}^{E-1}\nabla\widetilde{F}(\boldsymbol{\Omega}^{t,e}) = \sum_{e=0}^{E-1}\left(\boldsymbol{\Omega}^{t,e} - \boldsymbol{\Omega}^{t,e+1}\right)$$
$$= \boldsymbol{\Omega}^{t,0} - \boldsymbol{\Omega}^{t,e} = \boldsymbol{\Omega}^t - \boldsymbol{\Omega}^{t,e}. \tag{30}$$

we then rewrite the server update in (21) as follows

$$\boldsymbol{\Omega}^{t+1} = (I - \eta\lambda\kappa_R\mathcal{L})\left[\boldsymbol{\Omega}^t - \eta\sum_{e=0}^{E-1}\nabla\widetilde{F}(\boldsymbol{\Omega}^{t,e})\right] \tag{31}$$

$$= \boldsymbol{\Omega}^t - \eta\sum_{e=0}^{E-1}\nabla\widetilde{F}(\boldsymbol{\Omega}^{t,e}) - \eta\lambda\kappa_R\mathcal{L}\boldsymbol{\Omega}^t$$
$$+ \frac{\eta^2\lambda\kappa_R\mathcal{L}}{E}\sum_{e=0}^{E-1}\nabla\widetilde{F}(\boldsymbol{\Omega}^{t,e})$$

$$= \boldsymbol{\Omega}^t - \eta\left(\sum_{e=0}^{E-1}\nabla\widetilde{F}(\boldsymbol{\Omega}^{t,e}) - \lambda\kappa_R\mathcal{L}\boldsymbol{\Omega}^t\right.$$
$$\left.+ \frac{\eta\lambda\kappa_R\mathcal{L}}{E}\sum_{e=0}^{E-1}\nabla\widetilde{F}(\boldsymbol{\Omega}^{t,e})\right)$$

$$= \boldsymbol{\Omega}^t - \eta\left(\sum_{e=0}^{E-1}\nabla\widetilde{\mathcal{J}}(\boldsymbol{\Omega}^{t,e}) + \frac{\lambda\kappa_R\mathcal{L}}{E}\sum_{e=0}^{E-1}(\boldsymbol{\Omega}^{t,e} - \boldsymbol{\Omega}^t)\right.$$
$$\left.+ \frac{\eta\lambda\kappa_R\mathcal{L}}{E}\sum_{e=0}^{E-1}\nabla\widetilde{F}(\boldsymbol{\Omega}^{t,e})\right),$$

which finishes the proof. □

Lemma 4. Suppose that Assumptions 1 to 3 hold. We can derive the gradient bound of the server update in Lemma 3 as follows:

$$\mathbb{E}\|Z^t\|^2 \leq (1 + \eta^2\lambda^2\kappa_R^2\mathcal{L}^2)\left(6L_J^2\mathcal{E}^{(t)} + 6\mathbb{E}\|\nabla J(\boldsymbol{\Omega}^t)\|^2\right.$$
$$\left.+ \frac{3\sigma_F^2}{E}\right) + 3\lambda^2\kappa_R^2\|\mathcal{L}\|^2\mathcal{E}^{(t)} \tag{32}$$

Proof: Using Jensen's inequality, we have that

$$\mathbb{E}\|Z^t\|^2 \leq 3\mathbb{E}\left\|\frac{1}{E}\sum_{e=0}^{E-1}\nabla\widetilde{\mathcal{J}}(\boldsymbol{\Omega}^{t,e})\right\|^2 \tag{33}$$
$$+ 3\mathbb{E}\left\|\frac{\lambda\kappa_R\mathcal{L}}{E}\sum_{e=0}^{E-1}\mathcal{L}(\boldsymbol{\Omega}^{t,e} - \boldsymbol{\Omega}^t)\right\|^2$$
$$+ 3\mathbb{E}\left\|\frac{\eta\lambda\kappa_R\mathcal{L}}{E}\sum_{e=0}^{E-1}\nabla\widetilde{F}(\boldsymbol{\Omega}^{t,e})\right\|^2.$$

First, according to [45, Lemma 9] the definition of $\mathcal{E}^{(t)}$, we can bound the the first term in (34) as follows:

$$3\mathbb{E}\left\|\frac{1}{E}\sum_{e=0}^{E-1}\nabla\widetilde{\mathcal{J}}(\boldsymbol{\Omega}^{t,e})\right\|^2 \tag{34}$$
$$\leq \frac{3}{E}\sum_{e=0}^{E-1}\mathbb{E}\|\nabla J(\boldsymbol{\Omega}^{t,e})\|^2 + \frac{3\sigma_F^2}{E}$$
$$\leq \frac{3}{E}\sum_{e=0}^{E-1}\left(2L_J^2\mathbb{E}\|\boldsymbol{\Omega}^{t,e} - \boldsymbol{\Omega}^t\|^2\right.$$
$$\left.+ 2\mathbb{E}\|\nabla J(\boldsymbol{\Omega}^t)\|^2\right) + \frac{3\sigma_F^2}{E}$$
$$= 6L_J^2\mathcal{E}^{(t)} + 6\mathbb{E}\|\nabla J(\boldsymbol{\Omega}^t)\|^2 + \frac{3\sigma_F^2}{E}.$$

Next, we can bound the second term in (34) as follows:

$$3\mathbb{E}\left\|\frac{\lambda\kappa_R\mathcal{L}}{E}\sum_{e=0}^{E-1}\mathcal{L}\left(\mathbf{\Omega}^{t,e}-\mathbf{\Omega}^t\right)\right\|^2$$

$$=3\lambda^2\kappa_R^2\mathbb{E}\left\|\frac{1}{E}\sum_{e=0}^{E-1}\mathcal{L}\left(\mathbf{\Omega}^{t,e}-\mathbf{\Omega}^t\right)\right\|^2$$

$$\leq\frac{3\lambda^2\kappa_R^2}{E}\sum_{e=0}^{E-1}\mathbb{E}\left\|\mathcal{L}\left(\mathbf{\Omega}^{t,e}-\mathbf{\Omega}^t\right)\right\|^2$$

$$\leq\frac{3\lambda^2\kappa_R^2}{E}\|\mathcal{L}\|^2\sum_{e=0}^{E-1}\mathbb{E}\left\|\left(\mathbf{\Omega}^{t,e}-\mathbf{\Omega}^t\right)\right\|^2=3\lambda^2\kappa_R^2\|\mathcal{L}\|^2\mathcal{E}^{(t)}.$$
(35)

Then, proceeding as in (35), the third term in (34) can be bounded as follows:

$$3\mathbb{E}\left\|\frac{\eta\lambda\kappa_R\mathcal{L}}{E}\sum_{e=0}^{E-1}\nabla\widetilde{F}\left(\mathbf{\Omega}^{t,e}\right)\right\|^2$$

$$\leq 3\eta^2\lambda^2\kappa_R^2\mathcal{L}^2\mathbb{E}\left\|\frac{1}{E}\sum_{e=0}^{E-1}\nabla\widetilde{F}\left(\mathbf{\Omega}^{t,e}\right)\right\|^2$$

$$\leq\eta^2\lambda^2\kappa_R^2\mathcal{L}^2\left(6L_J^2\mathcal{E}^{(t)}+6\mathbb{E}\left\|\nabla J\left(\mathbf{\Omega}^t\right)\right\|^2+\frac{3\sigma_F^2}{E}\right).$$
(36)

Substituting (35), (35), and (36) into (34), the result of this Lemma in (32) can be obtained. This concludes the proof. $\square$

Lemma 5. Let $\mathcal{E}^t:=\frac{1}{E}\sum_{e=0}^{E-1}\mathbb{E}\left\|\mathbf{\Omega}^{t,e}-\mathbf{\Omega}^t\right\|^2$ be the drift caused by E local update steps at clients, where $\mathbb{E}$ is the expectation taken over all random sources and $\mathbf{\Omega}^t=\mathbf{\Omega}^{t,0}$. Suppose that Assumption 3 holds, we have

$$\mathcal{E}^t\leq 4\eta^2\mathbb{E}\left\|\nabla F\left(\mathbf{\Omega}^t\right)\right\|^2+2\eta^2\sigma_F^2E$$

Proof: By Assumption 3, using Lemmas 3(a) and 3, we derive that

$$\mathbb{E}\left\|\mathbf{\Omega}^{t,e}-\mathbf{\Omega}^t\right\|^2=\mathbb{E}\left\|\mathbf{\Omega}^{t,e-1}-\mathbf{\Omega}^t-\eta\nabla\widetilde{F}\left(\mathbf{\Omega}^{t,e-1}\right)\right\|^2$$

$$\leq\mathbb{E}\left\|\mathbf{\Omega}^{t,e-1}-\mathbf{\Omega}^t-\eta\nabla F\left(\mathbf{\Omega}^{t,e-1}\right)\right\|^2+\eta^2\sigma_F^2$$

$$\leq\left(1+\frac{1}{E}\right)\mathbb{E}\left\|\mathbf{\Omega}^{t,e-1}-\mathbf{\Omega}^t\right\|^2$$
$$+(1+E)\eta^2\mathbb{E}\left\|\nabla F\left(\mathbf{\Omega}^t\right)\right\|^2+\eta^2\sigma_F^2$$

$$\leq\left(1+\frac{1}{E}\right)\mathbb{E}\left\|\mathbf{\Omega}^{t,e-1}-\mathbf{\Omega}^t\right\|^2$$
$$+\frac{2\eta^2}{E}\mathbb{E}\left\|\nabla F\left(\mathbf{\Omega}^t\right)\right\|^2+\eta^2\sigma_F^2,\quad(37)$$

where the last inequality is due to the fact that $1+R\tau\leq R+R=2R$ since $R\geq 1$ and $\tau\leq 1$. Telescoping the last inequality yields

$$\mathbb{E}\left\|\mathbf{\Omega}^{t,e}-\mathbf{\Omega}^t\right\|^2$$
(38)
$$\leq\left(\frac{2\tilde{\eta}^2}{E}\mathbb{E}\left\|\nabla F\left(\mathbf{\Omega}^t\right)\right\|^2+\frac{\tilde{\eta}^2\sigma_F^2}{R^2}\right)\sum_{r=1}^{R-1}\left(1+\frac{1}{E}\right)^r.$$

Since $\sum_{j=0}^{m-1}x_j=\frac{x^m-1}{x-1}$ and $\left(1+\frac{x}{n}\right)^n\leq e^x,\forall x\in\mathbb{R},n\in\mathbb{N}$, we have $\sum_{e=0}^{E-1}\left(1+\frac{1}{E}\right)^r=\frac{\left(1+\frac{1}{E}\right)^E-1}{\left(1+\frac{1}{E}\right)-1}\leq(e-1)E\leq 2E$, and thus

$$\mathbb{E}\left\|\mathbf{\Omega}^{t,e}-\mathbf{\Omega}^t\right\|^2\leq 4\eta^2\mathbb{E}\left\|\nabla F\left(\mathbf{\Omega}^t\right)\right\|^2+2\eta^2\sigma_F^2E.\quad(39)$$

Averaging it over r, we get the conclusion. $\square$

A.2 Proof of Theorem 1

First, according to the result of Lemma 3, we can have

$$\mathbb{E}\left\|\mathbf{\Omega}^{t+1}-\mathbf{\Omega}^*\right\|^2=\mathbb{E}\left\|\mathbf{\Omega}^t-\mathbf{\Omega}^*\right\|^2+\tilde{\eta}^2\mathbb{E}\left\|Z^t\right\|^2$$
$$-2\tilde{\eta}\mathbb{E}\left\langle Z^t,\mathbf{\Omega}^t-\mathbf{\Omega}^*\right\rangle.\quad(40)$$

For the third term, we can obtain its bound from the Lemma 3 as follows:

$$-2\tilde{\eta}\mathbb{E}\left\langle Z^t,\mathbf{\Omega}^t-\mathbf{\Omega}^*\right\rangle$$
(41)

$$=\frac{2\tilde{\eta}}{E}\sum_{e=0}^{E-1}\mathbb{E}\left\langle\nabla J\left(\mathbf{\Omega}^{t,e}\right),\mathbf{\Omega}^*-\mathbf{\Omega}^t\right\rangle$$

$$+\frac{2\tilde{\eta}\lambda\kappa_R}{E}\sum_{e=0}^{E-1}\mathbb{E}\left\langle\mathcal{L}\left(\mathbf{\Omega}^{t,e}-\mathbf{\Omega}^t\right),\mathbf{\Omega}^t-\mathbf{\Omega}^*\right\rangle$$

$$+\frac{2\tilde{\eta}^2\lambda\kappa_R}{E}\sum_{e=0}^{E-1}\mathbb{E}\left\langle\mathcal{L}\nabla F\left(\mathbf{\Omega}^{t,e}\right),\mathbf{\Omega}^t-\mathbf{\Omega}^*\right\rangle$$

$$\leq\frac{2\tilde{\eta}}{E}\sum_{e=0}^{E-1}\left(\mathbb{E}\left[J\left(\mathbf{\Omega}^*\right)-J\left(\mathbf{\Omega}^t\right)\right]-\frac{\mu}{4}\mathbb{E}\left\|\mathbf{\Omega}^t-\mathbf{\Omega}^*\right\|^2\right.$$

$$+L_J\mathbb{E}\left\|\mathbf{\Omega}^{t,e}-\mathbf{\Omega}^t\right\|^2\bigg)$$

$$+\frac{2\tilde{\eta}\lambda\kappa_R}{E}\sum_{e=0}^{E-1}\left(\frac{m}{2}\|\mathcal{L}\|^2\mathbb{E}\left\|\mathbf{\Omega}^{t,e}-\mathbf{\Omega}^t\right\|^2\right.$$

$$+\frac{1}{2m}\mathbb{E}\left\|\mathbf{\Omega}^t-\mathbf{\Omega}^*\right\|^2\bigg)$$

$$+\frac{2\tilde{\eta}^2\lambda\kappa_R}{E}\sum_{e=0}^{E-1}\left(\frac{n}{2}\|\mathcal{L}\|^2\mathbb{E}\left\|\nabla F\left(\mathbf{\Omega}^{t,e}\right)\right\|^2\right.$$

$$+\frac{1}{2n}\mathbb{E}\left\|\mathbf{\Omega}^t-\mathbf{\Omega}^*\right\|^2\bigg),$$

where $m,n>0$ will be chosen later. Due to the smoothness of $F(\cdot)$, we have

$$\mathbb{E}\left\|\nabla F\left(\mathbf{\Omega}^{t,e}\right)\right\|^2\leq 2L^2\mathbb{E}\left\|\mathbf{\Omega}^{t,e}-\mathbf{\Omega}^t\right\|^2+2\mathbb{E}\left\|\nabla F\left(\mathbf{\Omega}^t\right)\right\|^2$$

$$\leq 2L^2\mathbb{E}\left\|\mathbf{\Omega}^{t,e}-\mathbf{\Omega}^t\right\|^2$$

$$+\frac{8L^2}{\mu}\mathbb{E}\left[J\left(\mathbf{\Omega}^t\right)-J\left(\mathbf{\Omega}^*\right)\right]+4\Gamma^L.(42)$$

Substituting (42) into (42) and setting $m=\frac{8\lambda}{\mu}$ and $n=\frac{8\tilde{\eta}\lambda}{\mu}$, we have

$$-2\tilde{\eta}\mathbb{E}\left\langle Z^t,\mathbf{\Omega}^t-\mathbf{\Omega}^*\right\rangle$$
(43)

$$\leq-\left(2\tilde{\eta}-\frac{64\tilde{\eta}^3\lambda^2\kappa_R^2\|\mathcal{L}\|^2L^2}{\mu^2}\right)\mathbb{E}\left[J\left(\mathbf{\Omega}^t\right)-J\left(\mathbf{\Omega}^*\right)\right]$$

$$-\frac{\tilde{\eta}\mu}{4}\mathbb{E}\left\|\mathbf{\Omega}^t-\mathbf{\Omega}^*\right\|^2+\frac{32\tilde{\eta}^3\lambda^2\kappa_R^2\|\mathcal{L}\|^2\Gamma^L}{\mu}$$

$$+\tilde{\eta}\left(2L_J+\frac{\lambda^2\kappa_R^2\|\mathcal{L}\|^2}{\mu}+\frac{16\tilde{\eta}^2\lambda^2\kappa_R^2\|\mathcal{L}\|^2L^2}{\mu}\right)\mathcal{E}^{(t)}.$$

Combining this with (44) and Lemma 3, we rewrite (40) as follow:

$$
\mathbb{E}\left\|\mathbf{\Omega}^{t+1}-\mathbf{\Omega}^{*}\right\|^{2} \tag{44}
$$
$$
\begin{aligned}
\leq & \left(1-\frac{\tilde{\eta}\mu}{4}\right)\mathbb{E}\left\|\mathbf{\Omega}^{t}-\mathbf{\Omega}^{*}\right\|^{2}+\tilde{\eta}p\mathcal{E}^{(t)}+6\tilde{\eta}^{2}\mathbb{E}\left\|\nabla J\left(\mathbf{\Omega}^{t}\right)\right\|^{2} \\
& -\left(2\tilde{\eta}-\frac{64\tilde{\eta}^{3}\lambda^{2}\kappa_{R}^{2}\|\mathcal{L}\|^{2}L^{2}}{\mu^{2}}\right)\mathbb{E}\left[J\left(\mathbf{\Omega}^{t}\right)-J\left(\mathbf{\Omega}^{*}\right)\right] \\
& +6\tilde{\eta}^{4}\lambda^{2}\kappa_{R}^{2}\|\mathcal{L}\|^{2}\mathbb{E}\left\|\nabla F\left(\mathbf{\Omega}^{t}\right)\right\|^{2}+\frac{32\tilde{\eta}^{3}\lambda^{2}\kappa_{R}^{2}\|\mathcal{L}\|^{2}\Gamma^{L}}{\mu} \\
& +\frac{3\tilde{\eta}^{2}\left(1+\tilde{\eta}^{2}\lambda^{2}\kappa_{R}^{2}\|\mathcal{L}\|^{2}\right)\sigma_{F}^{2}}{E},
\end{aligned}
$$

where $p=2L_{J}+\frac{8\lambda^{2}\kappa_{R}^{2}\|\mathcal{L}\|^{2}}{\mu}+\frac{64\beta^{2}}{\mu}+\frac{12L_{J}^{2}}{\lambda\kappa_{R}\|\mathcal{L}\|}+6\lambda\kappa_{R}\|\mathcal{L}\|+\frac{48L^{2}}{\lambda\kappa_{R}\|\mathcal{L}\|}$.

Using Lemma 5, we have

$$
\mathbb{E}\left\|\mathbf{\Omega}^{t+1}-\mathbf{\Omega}^{*}\right\|^{2} \tag{45}
$$
$$
\begin{aligned}
\leq & \left(1-\frac{\tilde{\eta}\mu}{4}\right)\mathbb{E}\left\|\mathbf{\Omega}^{t}-\mathbf{\Omega}^{*}\right\|^{2} \\
& -\left(2\tilde{\eta}-\frac{64\tilde{\eta}^{3}\lambda^{2}\kappa_{R}^{2}\|\mathcal{L}\|^{2}L^{2}}{\mu^{2}}\right)\mathbb{E}\left[J\left(\mathbf{\Omega}^{t}\right)-J\left(\mathbf{\Omega}^{*}\right)\right] \\
& +\left(4p\tilde{\eta}^{3}+6\tilde{\eta}^{4}\lambda^{2}\kappa_{R}^{2}\|\mathcal{L}\|^{2}\right)\left(4\frac{L^{2}}{\mu}\mathbb{E}\left[J(\mathbf{\Omega})-J\left(\mathbf{\Omega}^{*}\right)\right]\right. \\
& \left.+2\Gamma^{L}\right)+\frac{2p\tilde{\eta}^{3}\tau^{2}\sigma_{F}^{2}/B}{E}+12\tilde{\eta}^{2}L_{J}\mathbb{E}\left[J(\mathbf{\Omega})-J\left(\mathbf{\Omega}^{*}\right)\right] \\
& +\frac{32\tilde{\eta}^{3}\lambda^{2}\kappa_{R}^{2}\|\mathcal{L}\|^{2}\Gamma^{L}}{\mu}+\frac{3\tilde{\eta}^{2}\left(1+\tilde{\eta}^{2}\lambda^{2}\kappa_{R}^{2}\|\mathcal{L}\|^{2}\right)\sigma_{F}^{2}/B}{E} \\
\leq & \left(1-\frac{\tilde{\eta}\mu}{4}\right)\mathbb{E}\left\|\mathbf{\Omega}^{t}-\mathbf{\Omega}^{*}\right\|^{2}+\tilde{\eta}^{3}\underbrace{\frac{p\left(8E\Gamma^{L}+2\sigma_{F}^{2}/B\right)}{E}}_{C_{2}} \\
& -\left[2\tilde{\eta}-\tilde{\eta}^{2}q\right]\mathbb{E}\left[J\left(\mathbf{\Omega}^{t}\right)-J\left(\mathbf{\Omega}^{*}\right)\right] \\
& +\tilde{\eta}^{2}\underbrace{\frac{(64\lambda\kappa_{R}\|\mathcal{L}\|E+48)\Gamma^{L}+15\sigma_{F}^{2}/B}{\mu E}}_{C_{1}},
\end{aligned}
$$

where $q=\left(\frac{128\lambda\kappa_{R}\|\mathcal{L}\|L^{2}}{\mu^{2}}+12L_{J}+\frac{96L^{2}}{\mu}+\frac{32pL^{2}}{\mu\lambda\kappa_{R}\|\mathcal{L}\|}\right)$. Let $\mu\leq\frac{\tilde{\eta}_{1}}{E}$. Then $\tilde{\eta}\leq\tilde{\eta}_{1}=\min\left\{\frac{1}{q},\frac{2}{\lambda\kappa_{R}\|\mathcal{L}\|}\right\}\leq\frac{1}{q}$, which implies that $2\tilde{\eta}-\tilde{\eta}^{2}q\geq\tilde{\eta}$, and so

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{\Omega}^{t+1}-\mathbf{\Omega}^{*}\right\|^{2}\leq & \left(1-\frac{\tilde{\eta}\mu}{4}\right)\mathbb{E}\left\|\mathbf{\Omega}^{t}-\mathbf{\Omega}^{*}\right\|^{2} \\
& -\tilde{\eta}\mathbb{E}\left[J\left(\mathbf{\Omega}^{t}\right)-J\left(\mathbf{\Omega}^{*}\right)\right]+\tilde{\eta}^{3}C_{2}+\tilde{\eta}^{2}C_{1}
\end{aligned} \tag{46}
$$

Recalling that $\Delta^{(t)}=\left\|\mathbf{\Omega}^{t}-\mathbf{\Omega}^{*}\right\|^{2}$, rearranging the terms, and multiplying both sides of (66) with $\frac{\theta^{(t)}}{\tilde{\eta}\tau\Theta_{T}}$, where $\Theta_{T}=\sum_{t=0}^{T-1}\theta^{(t)}$, we obtain that

$$
\begin{aligned}
& \sum_{t=0}^{T-1}\frac{\theta^{(t)}\mathbb{E}\left[J\left(\mathbf{\Omega}^{t}\right)\right]}{\Theta_{T}}-J\left(\mathbf{\Omega}^{*}\right) \\
& \leq\sum_{t=0}^{T-1}\mathbb{E}\left[\left(1-\frac{\tilde{\eta}\mu}{4}\right)\frac{\theta^{(t)}\Delta^{(t)}}{\tilde{\eta}\tau\Theta_{T}}-\frac{\theta^{(t)}\Delta^{(t+1)}}{\tilde{\eta}\tau\Theta_{T}}\right] \\
& \quad+\mu^{2}\tau C_{2}+\tilde{\eta}C_{1} \\
& =\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{\theta^{(t-1)}\Delta^{(t)}-\theta^{(t)}\Delta^{(t+1)}}{\tilde{\eta}\tau\Theta_{T}}\right]+\mu^{2}\tau C_{2}+\tilde{\eta}C_{1} \\
& =\frac{1}{\tilde{\eta}\tau\Theta_{T}}\Delta^{(0)}-\frac{\theta^{(T-1)}}{\tilde{\eta}\tau\Theta_{T}}\mathbb{E}\Delta^{(T)}+\tilde{\eta}^{2}C_{2}+\tilde{\eta}C_{1} \\
& \leq\frac{1}{\tilde{\eta}\tau\Theta_{T}}\Delta^{(0)}+\tilde{\eta}^{2}C_{2}+\tilde{\eta}C_{1}.
\end{aligned} \tag{47}
$$

Here, (47) follows from the fact that $\left(1-\frac{\tilde{\eta}\mu}{4}\right)\theta^{(t)}=\theta^{(t-1)}$ due to $\theta^{(t)}=\left(1-\frac{\tilde{\eta}\mu}{4}\right)^{-(t+1)}$. Now, let $T\geq\frac{4E}{\tilde{\eta}_{1}\mu S}$. There is $\left(1-\frac{\tilde{\eta}\mu}{4}\right)^{T}\leq\exp\left(-\frac{\tilde{\eta}\mu T}{4}\right)\leq\exp(-1)\leq\frac{3}{4}$, and thus

$$
\Theta_{T}\geq\left(1-\frac{\tilde{\eta}\mu}{4}\right)^{-T}\frac{1}{\tilde{\eta}\mu}=\frac{\theta^{(T-1)}}{\tilde{\eta}\mu}. \tag{48}
$$

which yields $\frac{1}{\tilde{\eta}\Theta_{T}}\leq\frac{\mu}{\theta^{(T-1)}}\leq\mu e^{-\frac{\tilde{\eta}\mu T}{4}}$. Therefore, (47) can be rewritten as follows:

$$
\sum_{t=0}^{T-1}\frac{\theta^{(t)}\mathbb{E}\left[J\left(\mathbf{\Omega}^{t}\right)\right]}{\Theta_{T}}-J\left(\mathbf{\Omega}^{*}\right)\leq\mu\Delta^{(0)}e^{-\frac{\tilde{\eta}\mu T}{4}}+\tilde{\eta}^{2}C_{2}+\tilde{\eta}C_{1}, \tag{49}
$$

which together with the convexity of J implies that

$$
\begin{aligned}
\mathbb{E}\left[J\left(\widetilde{\mathbf{\Omega}}^{T}\right)-J\left(\mathbf{\Omega}^{*}\right)\right] & =\mathbb{E}\left[J\left(\sum_{t=0}^{T-1}\frac{\theta^{(t)}}{\Theta_{T}}\mathbf{\Omega}^{t}\right)\right]-J\left(\mathbf{\Omega}^{*}\right) \\
& \leq\mu\Delta^{(0)}e^{-\frac{\tilde{\eta}\mu T}{4}}+\tilde{\eta}^{2}C_{2}+\tilde{\eta}C_{1}.
\end{aligned} \tag{50}
$$

Using (49) -(50) and by the L-smoothness of $J(\cdot)$, we can easily obtain ,

$$
\begin{aligned}
\mathbb{E}\left[J\left(\mathbf{\Omega}^{T}\right)-J\left(\mathbf{\Omega}^{*}\right)\right] & \leq\frac{L}{2}\left(\mathbb{E}\left[\left\|\mathbf{\Omega}^{T}-\mathbf{\Omega}^{*}\right\|^{2}\right]\right) \\
& \leq\frac{L}{\mu}\left(\mu\Delta^{(0)}e^{-\frac{\tilde{\eta}\mu T}{4}}+\tilde{\eta}^{2}C_{2}+\tilde{\eta}C_{1}\right) \\
& =\frac{L}{\mu}\mathcal{O}\left(\mathbb{E}\left[J\left(\widetilde{\mathbf{\Omega}}^{T}\right)-J\left(\mathbf{\Omega}^{*}\right)\right]\right).
\end{aligned} \tag{51}
$$

Following the same approaches in [40], we consider the following cases.

- If $\tilde{\eta}_{1}\geq\widehat{\mu}:=\max\left\{\frac{4}{\mu T},\frac{4}{\mu T}\log\left(\frac{\mu^{2}\Delta^{(0)}T}{C_{1}}\right)\right\}$, then we choose $\mu=\widehat{\mu}$ and have

$$
\mathbb{E}\left[J\left(\widetilde{\mathbf{\Omega}}^{T}\right)-J\left(\mathbf{\Omega}^{*}\right)\right]\leq\widetilde{\mathcal{O}}\left(\frac{C_{2}}{\mu^{2}T^{2}}\right)+\widetilde{\mathcal{O}}\left(\frac{C_{1}}{\mu T}\right) \tag{52}
$$

- If $\frac{4}{\mu T}\leq\tilde{\eta}_{1}\leq\widehat{\mu}$, then we choose $\mu=\tilde{\eta}_{1}$ and have

$$
\begin{aligned}
\mathbb{E}\left[J\left(\widetilde{\mathbf{\Omega}}^{T}\right)-J\left(\mathbf{\Omega}^{*}\right)\right] & \leq\mathcal{O}\left(\alpha\Delta^{(0)}e^{-\frac{\tilde{\eta}_{1}\mu T}{4}}\right) \\
& +\widetilde{\mathcal{O}}\left(\frac{C_{2}}{\mu^{2}T^{2}}\right)+\widetilde{\mathcal{O}}\left(\frac{C_{1}}{\mu T}\right)
\end{aligned} \tag{53}
$$

By combining (51) and the above two cases,

$$\mathbb{E}\left[J\left(\mathbf{\Omega}^T\right) - J\left(\mathbf{\Omega}^*\right)\right] \tag{54}$$

$$\leq \mathcal{O}\left(\Delta^{(0)}e^{-\frac{\tilde{\eta}_1\mu^2 T}{4}} + \frac{(1+\mu T)(E\Gamma^L + \sigma_F^2/B)}{\mu^3 T^2 EK^L}\right),$$

where $\Delta^{(0)} = \|\mathbf{\Omega}^0 - \mathbf{\Omega}^*\|^2$. This concludes the proof.

APPENDIX B
PROOF OF THEOREM 2

In this section, we would like to bound the transfer objective $\mathbb{E}[F_{k''}(\boldsymbol{\omega}_{k''}) - F_{k''}(\boldsymbol{\omega}_{k''}^*)]$ in (6) to capture the above-performance gap between the transfer model obtained by our proposed transfer solution and the local optimal model. Note that the expected error bound of $\mathbb{E}[F_{k''}(\boldsymbol{\omega}_{k''}) - F_{k''}(\boldsymbol{\omega}_{k''}^*)]$ can be decomposed into the following form:

$$\underbrace{\mathbb{E}[F_{k''}(\boldsymbol{\omega}_{k''}) - F_{k''}(\boldsymbol{\omega}_{k''}^*)]}_{\text{transfer error}}$$

$$\leq \underbrace{\mathbb{E}[F_{k''}(\boldsymbol{\omega}_{k''}) - \frac{1}{K^L}\sum_{k'\in\mathcal{K}^L}F_{k'}(\boldsymbol{\omega}_{k'})]}_{\text{generalization error}}$$

$$+ \underbrace{\mathbb{E}[\frac{1}{K^L}\sum_{k'\in\mathcal{K}^L}F_{k'}(\boldsymbol{\omega}_{k'})] - \frac{1}{K^L}\sum_{k'\in\mathcal{K}^L}F_{k'}(\boldsymbol{\omega}_{k'}^*)}_{\text{training error}} \tag{55}$$

$$+ \underbrace{\frac{1}{K^L}\sum_{k'\in\mathcal{K}^L}F_{k'}(\boldsymbol{\omega}_{k'}^*) - F_{k''}(\boldsymbol{\omega}_{k''}^*)}_{\text{transfer gap}}.$$

Hence, to bound the expected transfer error, we should bound the expectation of training and generalization errors. To begin with, we first introduce the following lemmas.

B.1 Key Lemmas of Theorem 2

Lemma 6. Suppose Assumptions 5 holds. According to the proposed model transfer solution, we can bound the generalization error term in (55) as follows:

$$\mathbb{E}[F_{k''}(\boldsymbol{\omega}_{k''}) - \frac{1}{K^L}\sum_{k'\in\mathcal{K}^L}F_{k'}(\boldsymbol{\omega}_{k'})]$$

$$\leq M\sum_{k'\in\mathcal{K}^L}\|\xi_{k'',k'}\mathcal{P}_{k''} - \frac{\mathcal{P}_{k'}}{K^L}\|_{TV}. \tag{56}$$

Proof: Due to $\boldsymbol{\omega}_{k''} = \sum_{k'\in\mathcal{K}^L}\xi_{k'',k'}\boldsymbol{\omega}_{k'}$, using Jensen's inequality, we have $F_{k''}(\boldsymbol{\omega}_{k''}) \leq \sum_{k'\in\mathcal{K}^L}\xi_{k'',k'}F_{k''}(\boldsymbol{\omega}_{k'})$. Therefore, we can derive the following result:

$$\mathbb{E}[F_{k''}(\boldsymbol{\omega}_{k''}) - \frac{1}{K^L}\sum_{k'\in\mathcal{K}^L}F_{k'}(\boldsymbol{\omega}_{k'})]$$

$$\leq \sum_{k'\in\mathcal{K}^L}\mathbb{E}[\xi_{k'',k'}F_{k''}(\boldsymbol{\omega}_{k'}) - \frac{1}{K^L}F_{k'}(\boldsymbol{\omega}_{k'})]$$

$$\leq M\sum_{k'\in\mathcal{K}^L}\|\xi_{k'',k'}\mathcal{P}_{k''} - \frac{\mathcal{P}_{k'}}{K^L}\|_{TV}. \tag{57}$$

The last inequality in (57) can be obtained by using Definition 2. This concludes the proof. □

Lemma 7. Let $\mathbb{E}[\mathcal{J}(\mathbf{\Omega}^T) - \mathcal{J}(\mathbf{\Omega}^*)] \leq \varepsilon$ hold. When choosing $R(\|\boldsymbol{\omega}_{k'} - \boldsymbol{\omega}_{j'}\|^2) = 1 - e^{-\|\boldsymbol{\omega}_{k'} - \boldsymbol{\omega}_{j'}\|^2/\sigma_R}$ with parameter σ_R, the training error term in (55) can be bounded as follows:

$$\mathbb{E}[\frac{1}{K^L}\sum_{k'\in\mathcal{K}^L}F_{k'}(\boldsymbol{\omega}_{k'})] - \frac{1}{K^L}\sum_{k'\in\mathcal{K}^L}F_{k'}(\boldsymbol{\omega}_{k'}^*) \leq \frac{\varepsilon}{K^L} + \lambda. \tag{58}$$

Proof: According to the Theorem 1, we have $\mathbb{E}\left[J\left(\mathbf{\Omega}^T\right) - J\left(\mathbf{\Omega}^*\right)\right] = \mathbb{E}\left[F\left(\mathbf{\Omega}^T\right) - F\left(\mathbf{\Omega}^*\right)\right] + \lambda\left(\mathcal{R}(\mathbf{\Omega}) - \mathcal{R}(\mathbf{\Omega}^*)\right) \leq \varepsilon$. When choosing $R(\|\boldsymbol{\omega}_{k'} - \boldsymbol{\omega}_{j'}\|^2) = 1 - e^{-\|\boldsymbol{\omega}_{k'} - \boldsymbol{\omega}_{j'}\|^2/\sigma_R}$, we can derive that

$$\frac{1}{K^L}\mathbb{E}\left[\mathcal{F}\left(\mathbf{\Omega}^T\right) - \mathcal{F}\left(\mathbf{\Omega}^*\right)\right] \tag{59}$$

$$= \frac{1}{K^L}\mathbb{E}\left[\mathcal{J}\left(\mathbf{\Omega}^T\right) - \mathcal{J}\left(\mathbf{\Omega}^*\right)\right] + \frac{\lambda}{K^L}\left(\mathcal{R}(\mathbf{\Omega}) - \mathcal{R}(\mathbf{\Omega}^*)\right)$$

$$< \frac{\varepsilon}{K^L} + \lambda.$$

The inequality in (60) holds due to $0 < R(\|\boldsymbol{\omega}_{k'} - \boldsymbol{\omega}_{j'}\|^2) < 1$ in this case. This concludes the proof. □

B.2 Proof of Theorem 2

Proof: By combining Lemma 6 and Lemma 7, we can rewrite (55) as follows:

$$\mathbb{E}[F_{k''}(\boldsymbol{\omega}_{k''}) - F_{k''}(\boldsymbol{\omega}_{k''}^*)] \tag{60}$$

$$\leq M\sum_{k'\in\mathcal{K}^L}\left\|\xi_{k'',k'}\mathcal{P}_{k''} - \frac{\mathcal{P}_{k'}}{K^L}\right\|_{TV} + \frac{\epsilon + (\lambda + \Gamma^N)K^L}{K^L},$$

where $\Gamma^N = \frac{1}{K^L}\sum_{k'\in\mathcal{K}^L}F_{k'}(\boldsymbol{\omega}_{k'}^*) - F_{k''}(\boldsymbol{\omega}_{k''}^*)$. This concludes the proof. □