AutoFi: Toward Automatic Wi-Fi Human Sensing via Geometric Self-Supervised Learning
Jianfei Yang, Member, IEEE, Xinyan Chen, Han Zou, Dazhuo Wang, and Lihua Xie, Fellow, IEEE

Abstract—Wi-Fi sensing technology has shown superiority in smart homes among various sensors for its cost-effective and privacy-preserving merits. It is empowered by channel state information (CSI) extracted from Wi-Fi signals and advanced machine learning models to analyze motion patterns in CSI. Many learning-based models have been proposed for kinds of applications, but they severely suffer from environmental dependency. Though domain adaptation methods have been proposed to tackle this issue, it is not practical to collect high-quality, well-segmented, and balanced CSI samples in a new environment for adaptation algorithms, but randomly captured CSI samples can be easily collected. In this article, we first explore how to learn a robust model from these low-quality CSI samples, and propose AutoFi, an annotation-efficient Wi-Fi sensing model based on a novel geometric self-supervised learning algorithm. The AutoFi fully utilizes unlabeled low-quality CSI samples that are captured randomly, and then transfers the knowledge to specific tasks defined by users, which is the first work to achieve cross-task transfer in Wi-Fi sensing. The AutoFi is implemented on a pair of Atheros Wi-Fi APs for evaluation. The AutoFi transfers knowledge from randomly collected CSI samples into human gait recognition and achieves state-of-the-art performance. Furthermore, we simulate cross-task transfer using public data sets to further demonstrate its capacity for cross-task learning. For the UT-HAR and Widar data sets, the AutoFi achieves satisfactory results on activity recognition and gesture recognition without any prior training. We believe that AutoFi takes a huge step toward automatic Wi-Fi sensing without any developer engagement. Our codes have been included in https://github.com/xyanchen/Wi-Fi-CSI-Sensing-Benchmark.

Index Terms—Activity recognition, channel state information (CSI), deep learning, gait recognition, self-supervised learning, Wi-Fi sensing.

I. INTRODUCTION

WITH the increasing demands of Internet access, Wi-Fi infrastructures have been ubiquitous and many mobile devices are equipped with Wi-Fi modules. Multiple-input multiple-output (MIMO) with orthogonal frequency-division multiplexing (OFDM) was innovated for the higher requirement of data traffic in wireless communications [1]. Along with very high-spectral efficiency, MIMO provides the channel state information (CSI) for antenna pairs between receiver and transmitter devices. The CSI data records the propagation quality of multipath wireless signals in specific environments, and therefore it enables Wi-Fi-based radar technology [2], [3]. Wi-Fi-based radar can sense human motions by extracting CSI patterns by signal processing [4] or data-driven models [5], which has empowered many applications at smart homes, including occupancy estimation [6], activity recognition [7], gesture recognition [8], [9], human identification [10], human pose estimation [11], and vital sign detection [4].

Wi-Fi sensing methods can be categorized into model-based methods and learning-based methods that serve for different applications. Model-based methods formulate the Wi-Fi signals and its environment by physical models, such as the Fresnel zone [4]. For periodic human motions or simple activities, such as respiration and falling down [12], [13], the model-based methods are accurate and robust to environmental variations. However, it is hard to build physical models for complicated activities or compound motions. To deal with it, learning-based models are developed as deep learning models show a stronger capacity of extracting and modeling CSI patterns of complex gestures [14]. Nevertheless, the performance and generalization ability of data-driven models depend on the scale and variety of training samples, since the data collection and annotation process is usually time-consuming and labor-intensive. The model-based methods have achieved remarkable performance and robustness, so we mainly study the gap between the current learning-based methods and real-world challenging and practical applications.

Generally speaking, learning-based methods rely on statistical or deep learning models that map the CSI data to the label space in terms of specific tasks, such as the identity for human identification or gesture category for gesture recognition [5]. It is noteworthy that the success of deep learning models for visual recognition is dependent on the scale of the data set, e.g., the large-scale ImageNet [15], but such scale of data set does not exist in Wi-Fi sensing. The reason lies in the difficulty of collecting CSI samples by thousands of volunteers under thousands of circumstances. Recent work contributes to a bigger data set, such as Widar [16], but its scale is still below the ImageNet. Without sufficient data, learning-based models may fail in a new environment. Then many works commence to explore domain adaptation to deal with cross-environment problems, such as EL system [17] and WiADG [9]. These works are based on the domain adaptation methods that adapt...
the model to a new environment by minimizing the distribution discrepancy of the feature spaces between training and testing scenarios, which significantly improves the performance in the new environment [18]. However, to enable the domain adaptation methods, we need to collect a great number of high-quality CSI samples in the new environment, though in an unlabeled manner, but the data should be large-scale and balanced to all categories. Such assumption is naturally hard to achieve for real-world applications where users still need to do laborious data collection.

To bridge the gap between learning-based models and realistic Wi-Fi sensing, we study how deep models can work in an automatic data-efficient manner in this article. In realistic Wi-Fi scenarios, two kinds of data are accessible. First, CSI samples of human daily activities can be obtained without the activity labels and the segmentation of activities in Widar [16]. This can be simply achieved by setting a variation threshold of CSI streams, which offers massive unlabeled low-quality CSI samples. Second, a few number of labeled data can be collected with the cooperation from the user for calibration purpose, which is similar to the existing mobile phone security system setup of face and fingerprint recognition. If these easily-collected data can be leveraged for learning-based models, then it is not necessary to train a model in advance and conduct the domain adaptation process. The whole model learning process is therefore automatic without manual data collection and annotations, and the system can be initiated by users easily.

To this end, we propose an annotation-efficient Wi-Fi sensing system, namely, AutoFi, which learns new environmental settings in a self-driven fashion. It is an automatic Wi-Fi representation learning framework that helps achieve automatic Wi-Fi human sensing with very few manual annotations. As shown in Fig. 1, after deploying the AutoFi in a new environment, AutoFi first collects randomly-segmented and randomly-distributed CSI samples for any human actions. These samples could be persons passing by or various daily activities that are easy to acquire. Then, the self-supervised learning module enables the AutoFi to learn CSI patterns in an unsupervised manner, i.e., without the engagement of any labels. After self-supervised learning, the model has been initiated well with new environments learned. Then, we can conduct few-shot learning by calibrating several high-quality samples from users. It is worth noting that the task and the gesture categories can be totally customized by users, no matter whether the new defined gestures have been seen or not. It is the first work that achieves cross-task transfer in Wi-Fi sensing. The AutoFi learns how to extract robust features from environmental CSI samples, and contributes to customized functions. Extensive experiments are conducted in the real world and public data sets to demonstrate the effectiveness of our method.

The contributions are summarized as follows.

1) We analyze the main gaps between learning-based methods and practical Wi-Fi sensing, and propose the AutoFi to deal with it.

2) In AutoFi, we propose a novel self-supervised learning framework based on prevailing contrastive learning and mutual information and further enhance its transferability by developing a novel geometric structural loss, which helps AutoFi to enable various downstream tasks.

3) The AutoFi achieves the cross-task transfer for Wi-Fi sensing. To the best of our knowledge, it is the first work that achieves automatic Wi-Fi sensing in new environments without any prior data collection.

4) The AutoFi system is implemented in the real world to validate its robustness. We also simulate the AutoFi using public data sets, e.g., Widar and UT-HAR, and the results are also superior to existing domain adaptive systems.

II. RELATED WORKS

A. Wi-Fi-Based Passive Human Sensing

Recently, Wi-Fi-based passive radar is appealing in smart homes due to its low cost and high granularity. Compared to visual sensing [19], Wi-Fi sensing is privacy-preserving and illumination-robust. Wi-Fi sensing relies on CSI that is extracted from specific Wi-Fi chips, such as Intel 5300 NIC [20] and Atheros NIC [21]. The number of subcarriers and antennas determines the resolution of the CSI data. The Intel 5300 NIC tool can extract 30 subcarriers of CSI from each pair of antennas with a 20-MHz bandwidth, while the Atheros tool can take out 114 subcarriers of CSI with 40 MHz. The CSI data records the surrounding objects or motions that affect the multipath propagation of wireless signals. This process can be depicted by some physical models, such as Fresnel zone [4]. Relying on model analytics and signal processing, Wi-Fi passive radar achieves high performance on detecting periodic motions and specific human activities. The signal tendency index (STI) is developed to identify the occupancy situation [22]. Want et al. [4] proposed a respiration detection system and investigates the effect of user location and orientation, which is very useful in healthcare. Currently, Wi-Fi sensing has widespread applications, including occupancy estimation [6], [22], [23], activity recognition [7], [11], [24], [25], [26], [27], [28], gesture recognition [8], [9], [29], human
identification [10], [30], [31], human pose estimation [11], and vital sign detection [4], [32].

B. Learning-Based Methods for Wi-Fi Sensing

However, for more complex human gestures or even customized activities by users, machine learning models contribute to better capacity to recognize them. Wang et al. [33] first proposed a human activity recognition system by statistical features (e.g., mean and peak) and traditional classifiers. Then the E-eyes system is developed to achieve better performance by dividing human activities into in-place and dynamic ones [34]. The FreeCount system leverages a feature selection scheme based on information theory to conduct people counting [23]. These early-stage works show good performance on normal activities, such as walking and sitting, but they cannot identify fine-grained subtle gestures. To enhance the model capacity for these gesture recognition, deep learning models are introduced. Zou et al. [5] proposed the DeepSense that learns spatial-temporal features based on the convolutional neural network (CNN) and recurrent neural network. Yang et al. [35] proposed the EfficientFi that realizes the large-scale Wi-Fi sensing models by learning-based CSI compression. SecureSense is proposed to deal with the adversarial attacks by learning prediction consistency [36]. Chen et al. [37] proposed a bi-directional LSTM for activity recognition. These machine learning and deep learning methods show great performance in a single environment, but cannot generalize well to a new environment. To address this issue, the adversarial domain adaptation methods transfer knowledge from a source domain to a new target domain using only unlabeled examples [9]. Then, domain adaptation [18] is a prevailing method for cross-environment Wi-Fi sensing applications, such as TransferSense [38]. Nevertheless, it is noted that we still need high-quality CSI samples that have same categories, balanced label distribution and well-segmented actions in the unlabeled target domain [39], which requires users to engage and thus is still cumbersome. Another solution is to generate target-like samples by the generative adversarial network, but this also demands a number of high-quality data [26]. Our proposed AutoFi deals with this problem by learning randomly-segmented and randomly-distributed samples for downstream tasks, and hence it achieves automatic learning models for Wi-Fi sensing in the real world.

C. Self-Supervised Learning and Few-Shot Learning

As the AutoFi consists of two phases based on self-supervised learning and few-shot learning, we also review some recent progress on these perspectives. Self-supervised learning is a promising method to learn feature representations in an unsupervised manner [40]. Previous self-supervised methods are designed for unsupervised visual feature learning and they mainly rely on designing handcrafted auxiliary tasks, such as context prediction [41] and rotation prediction [42]. They achieve good performance but the handcrafted tasks limit the generalization ability of models. Then, the contrastive methods come into existence [43], which learns features from multiple views of samples via metric learning. SimCLR proposes to minimize the cosine similarity between views of same samples and maximize the similarity between those of different samples [44]. Then, the BYOL [45] first abandons the negative samples and adopt asymmetric architecture to mitigate the collapsed solution. Maximizing mutual information for representation learning is also prevailing, such as deep InfoMax [46] and TWIST [47]. Though self-supervised learning helps to generate a discriminative feature space, it does not contain any supervision tasks. To enable real-world applications, we further consider a data-efficient learning scheme: few-shot learning. Few-shot learning aims to conduct classification or regression by learning only several samples, or even one sample (i.e., one-shot learning) [48]. It is highly related to metric learning that is widely applied to face recognition [49], where triplet loss is utilized to cluster the samples from the same category and separate the samples from different categories. Yang et al. proposed to leverage few-shot learning for Wi-Fi-based gesture recognition. However, in few-shot learning in a new environment, we still need to initialize the model parameters using labeled training data collected in another environment, and this may lead to a domain shift that hinders the model performance. In the AutoFi, we enable the model to learn the environment by itself, and then utilize few-shot learning for gesture recognition.

III. Method

A. Overview

The objective of the AutoFi design is to enable learning-based Wi-Fi sensing by minimizing manual efforts. As shown in Fig. 2, the AutoFi is composed of two modules: 1) a geometric self-supervised (GSS) learning module and 2) a few-shot calibration (FSC) module. In the self-supervised learning module, the randomly-collected CSI data is processed by an augmentation $A_{\epsilon}$ to generate two random views, and these two views are fed into the feature extractors $E_{\theta_1}$ and $E_{\theta_2}$ and the nonlinear functions $G_{\phi_1}$ and $G_{\phi_2}$ to produce two distributions. The GSS loss enforces these two prediction distributions to be consistent, which does not require any annotation. Then, the well-trained feature extractors $E_{\theta_1}$ and $E_{\theta_2}$ can be transferred to the FSC module. Users only need to calibrate some gestures for several times to enable the recognition system, which allows users to define customized gestures or tasks. For the few-shot training, we use the prototypical network as the backbone [50].

B. Geometric Self-Supervised Learning Module

The GSS learning module aims to learn CSI representations in an unsupervised manner. Prevailing self-supervised learning methods employ handcrafted auxiliary tasks or contrastive learning [40]. In our scenarios, the downstream tasks can be quite different from the training samples that are randomly collected, and thus requires better transferability and generalization ability, which motivates us to design the GSS based on contrastive learning due to its stronger generalization capability [45]. The GSS modules consists of an augmentation module $A_{\epsilon}$ with a hyper-parameter $\epsilon$, the feature extractors
by adding a Gaussian noise \( \text{of the CSI data, we augment the input sample } \mathbf{E} \) CSI samples, and an FSC module that enables users to easily enable the recognition services. The feature extractor \( \mathbf{E} \) the FSC is initialized by the self-supervised module.

\( \mathbf{E}_1 \) and \( \mathbf{E}_2 \) parameterized by \( \theta_1 \) and \( \theta_2 \), respectively, and the nonlinear functions \( \mathbf{G}_{\phi_1} \) and \( \mathbf{G}_{\phi_2} \) parameterized by \( \phi_1 \) and \( \phi_2 \), respectively. The feature extractors are normally CNNs and the nonlinear functions are just multilayer perceptrons (MLPs). The input data is the randomly-collected unlabeled CSI samples \( \{\mathbf{x}'_i\}_{i=1}^N \). Each CSI sample is a matrix such that \( \mathbf{x}'_i \in \mathbb{R}^{S \times T} \), where \( S \) denotes the number of subcarriers and \( T \) denotes the time duration.

**Multiview Generation:** First, we input the samples to the augmentation module \( \mathbf{A}_e \). The augmentation module aims to generate two views for self-supervised learning. The two views should be meaningful but randomly augmented, such as the random cropping for images. For CSI data, previous research shows that the noises on subcarriers can be modeled as the Gaussian noise [27]. Hence, without breaking the intrinsic information of the CSI data, we augment the input sample by adding a Gaussian noise \( \zeta \sim \mathcal{N}(\mu, \sigma^2) \)

\[
\mathbf{A}_e(\mathbf{x}') = \mathbf{x}' + \epsilon \zeta
\]

where \( \epsilon \) is the weight of the noise. We can generate two views \( \mathbf{x}'_1, \mathbf{x}'_2 \) by \( \mathbf{A}_e(\mathbf{x}') \).

The next step is to extract features by \( \mathbf{E}_{\theta_1} \). Here, we just leverage a series of convolutional layers for \( \mathbf{E}_{\theta_1} \) as successfully used in many previous works [5]. Then, the feature embeddings are generated, but this feature space is what we aim to do classification in the few-shot learning. For self-supervised learning, we need to separate the feature space by a nonlinear function \( \mathbf{G}_{\phi_1} \). The bottleneck layer \( \mathbf{G}_{\phi_1} \) ensures that the self-supervised learning will not affect the feature learning, as discovered in [45]. After \( \mathbf{E}_{\theta_1} \) and \( \mathbf{G}_{\phi_1} \), the feature distributions of the first view are calculated by

\[
P(\mathbf{x}'_i^1) = \mathbf{G}_{\phi_1}(\mathbf{E}_{\theta_1}(\mathbf{x}'_i)).
\]

The second view is processed by \( \mathbf{E}_{\theta_2} \) and \( \mathbf{G}_{\phi_2} \), in the same way. In this fashion, \( P(\mathbf{x}'_i^1) \) and \( P(\mathbf{x}'_i^1) \) are obtained.

**Probability Consistency:** How to design the unsupervised loss is the key of the GSS module. We propose a novel learning objective that first incorporates geometric structures for unsupervised learning, which can benefit the downstream few-shot task. In contrastive learning, the normal objective is to force the predictions of different views to be consistent. To this end, the probability consistency loss is formulated as

\[
L_p = \frac{1}{2B} \sum_{i=1}^B \left( D_{\text{KL}}(P_1^1 || P_2^1) + D_{\text{KL}}(P_2^1 || P_1^1) \right)
\]

where \( D_{\text{KL}}(\cdot||\cdot) \) denotes the Kullback–Leibler divergence of the two distributions. Since the KL divergence is an asymmetric measure of distributions, we use dual forms to make it symmetric. By the consistency loss, the model learns to perform consistently on two views in terms of the prediction probabilities.

**Mutual Information:** In our scenario, we require the feature extractor to have the transferability for downstream tasks. To this end, we aim to maximize the mutual information between CSI samples and the feature space for better transferability. From the information theory, the mutual information between the prediction distributions and the input space should be maximized. The mutual information between a random variable \( X \) and its predicted label \( Y \) is formulated by

\[
I(X, Y) = H(Y) - H(Y|X)
\]

where \( H(\cdot) \) is the information entropy. Increasing \( H(Y) \) drives the model to predict uniform distributions among classes while decreasing \( H(Y|X) \) drives the model confidence of its predictions. However, the mutual information cannot be calculated directly, and therefore we aim to maximize its approximation by

\[
L_m = h(\mathbb{E}_{x_i \in \mathcal{B}^1} p^1) + \mathbb{E}_{x_i \in \mathcal{B}} h(p^1)
\]

where \( \mathcal{B} \) is a batch of samples and \( h(p) = -\sum_i p_i \log p_i \) is the conditional entropy. \( L_m \) operates on both \( P_1 \) and \( P_2 \) for all samples. Mutual information loss is widely used in semisupervised learning and domain adaptation [51].

**Geometric Consistency:** For our system, apart from learning discriminative features from unlabeled CSI samples, we further require the AutoFi to empower recognition capacity via few-shot learning. Nevertheless, former self-supervised learning may not be tailored for this purpose. They mostly rely on probability consistency and information maximization that...
enable a discriminative feature space but do not consider the downstream few-shot tasks. To deal with this problem, we propose a novel geometric loss in the GSS module. The rational behind this stems from the feature space of few-shot learning. The few-shot learning is highly related to metric learning and prototypical networks [48], [50] which leverage the cluster of each category and their geometric relationship. With tight clusters and meaningful geometry, the test sample can be predicted by retrieving the category of the most similar sample or applying k-nearest neighbors strategy in the feature space. In our scenarios, traditional self-supervised learning frameworks fail to capture geometry while classic few-shot learning frameworks cannot work well due to the lack of labels. To utilize the geometry among unlabeled samples, we propose a geometric structural loss that forces the geometry of two batches of views to be consistent. The geometry of a batch of samples can be generated by the relationship of neighbors. For a sample \( x^i \) with distribution \( P^i \), its geometric embedding \( Q^i \) can be formulated as

\[
q_{ij} = \frac{K(P^j, P^i)}{\sum_{m=1, m \neq j}^B K(p_m, P^j)}
\]

(6)

where \( q_{ij} \) denotes the \( j \)th position of \( Q^i \), and \( K(\cdot, \cdot) \) is a similarity function. Here, we choose the cosine similarity as

\[
K(a, b) = \frac{1}{2} \left( \frac{a^T b}{\|a\|_2 \|b\|_2} + 1 \right).
\]

(7)

Note that the geometric embedding \( Q^i \) represents the relationship between \( x^i \) and all neighbors in the feature space. Then, we train the model to generate a consistent geometry on two views by applying the KL divergence

\[
D_{KL}(Q^i || Q^j).
\]

(8)

The geometric structural loss helps the model learn geometry of CSI samples and further learn the feature space in terms of metrics. In this manner, the GSS module can enhance the subsequent few-shot learning module.

The total objective of the loss is defined as

\[
L = L_p + \lambda L_m + \gamma L_g
\]

(9)

where \( \lambda \) and \( \gamma \) are two hyper-parameters that balance multiple objectives for better convergence. In self-supervised learning, as long as they have similar magnitudes, the convergence can be achieved easily.

C. Few-Shot Calibration Module

After the GSS module, we transfer the feature extractors \( E_{\theta_1} \) and \( E_{\theta_2} \) to the FSC module, and reuse it to train a classifier for few-shot learning. Note that the two feature extractors are very similar, so either one can be used in FSC, denoted as \( E_{\theta} \). Users only need to collect several samples to setup the AutoFi. The labeled samples are denoted as \( \{x^i, y^i\}_{i=1}^M \), where \( M \) is the number of labeled samples. The feature embedding can be obtained by feeding samples into the feature extractor \( E_{\theta} \), and a classifier \( F_{\psi} \) maps the feature to its labels. In FSC, we first minimize the standard cross-entropy loss

\[
L_c = -E_{(x, y)} \sum_k \mathbb{1}[y = k] \log (F_{\psi}(E_{\theta}(x^i)))
\]

(10)

where \( \mathbb{1}[y = k] \) means a 0-1 function that outputs 1 for the correct category \( k \). Then to better cluster the same-class samples, we calculate the prototypes of each class as \( c_k \), and draw the same-class samples together by minimizing the log-probability

\[
L_f = -\log p_{\theta, \psi}(y = k|x)
\]

(11)

where \( p_{\theta, \psi}(y = k|x) \) is constructed by the distance between the sample \( x^i \) and its correct class center, formulated as

\[
p_{\theta, \psi}(y = k|x^i) = \frac{\exp(-d(F_{\psi}(E_{\theta}(x^i)), c_k))}{\sum_{k'} \exp(-d(F_{\psi}(E_{\theta}(x^i)), c_{k'}))}
\]

(12)

where \( k' \) denotes all categories. Note that the gesture or activity category and even the recognition task can be customized by users. The FSC is a normal few-shot learning scheme motivated by a prototypical network [50]. Whereas, after the feature extractor learns the randomly-collected samples in the GSS, it is found that the convergence of the FSC module can be easily achieved and the performance is boosted. In this manner, the AutoFi can quickly adapt to any environment automatically, and users input can enable the AutoFi to perform many downstream tasks without cumbersome data collection and model training. The whole algorithm is illustrated in Algorithm 1.

### IV. Experiments

A. Setup

Evaluation Scenarios and Criterion: We evaluate the AutoFi on different Wi-Fi platforms and CSI data. First, the AutoFi is implemented on a real-world IoT system for evaluation, demonstrating the main novelty of the AutoFi—to learn the environment by self-supervised learning and perform downstream tasks by few shots. The real-time system is based on the Atheros CSI tool and fine-grained CSI data [21]. Then, we evaluate the effectiveness of the AutoFi using the UTHAR data set, which leverages Intel 5300 NIC with a sampled number (30) of CSI subcarriers [52]. The third experiment is

Algorithm 1: Automatic Wi-Fi Sensing Setup

**Step 1: Train the GSS module**

**Module:** the feature extractors \( E_{\theta_1}, E_{\theta_2} \), the non-linear functions \( G_{\phi_1}, G_{\phi_2} \).

**Input:** unlabeled CSI data \( \{x^i\}_{i=1}^N \)

**BEGIN:**

1. **while** epoch < total epoch **do**

2. Augment samples by \( A_k(x^i) = x^i + \epsilon \xi \)

3. Obtain feature probabilities of views via \( P(x^i) = G_{\phi_1}(E_{\theta_1}(A_k(x^i))) \)

4. Update \( \theta_1, \theta_2, \phi_1, \phi_2 \) by minimizing \( L_p + \lambda L_m + \gamma L_g \)

5. **end while**

**Output:** the model parameters \( \theta_1, \theta_2 \).

**Step 2: Train the FSC module**

**Module:** the classifier \( F_{\psi} \).

**Input:** a few labeled samples \( \{x^i, y^i\}_{i=1}^M \)

**BEGIN:**

6. Initialize \( E_\theta \) by either \( \theta_1 \) or \( \theta_2 \)

7. **while** epoch < total epoch **do**

8. Update \( \theta, F_{\psi} \) by minimizing \( L_c + L_f \)

9. **end while**

**Output:** the model parameters \( \theta, \psi \).

**END.**
TABLE I
NETWORK ARCHITECTURE USED IN THE AUTOFI EXPERIMENTS. FOR CONV A × (H, W), WHERE A DENOTES THE CHANNEL NUMBER, AND (H, W) REPRESENTS THE HEIGHT AND WIDTH OF THE OPERATING KERNEL. THIS APPLIES TO ALL CONVOLUTION (CONV) AND MAX-POOLING (MAX-POOL) LAYERS

| Layer Index | Feature Extractor $E_f$ | Classifier $F_f$ |
|-------------|------------------------|-----------------|
| input       | CSI data: 3 × 114 × 500 |                 |
| 1           | Conv 32 × (15, 23), stride 9, ReLU | 128 dense |
| 2           | Conv 32 × (3, 7), stride 1, ReLU | 6 dense, softmax |
| 3           | Max-pool (1, 2), stride (1, 2) |                 |
| 4           | Conv 64 × (3, 7), stride 1, ReLU |                 |
| 5           | Conv 96 × (3, 7), stride 1, ReLU |                 |
| 6           | Max-pool (1, 2), stride (1, 2) |                 |

conducted on a large dataset, Widar [16]. Due to the different collection scenario, it is used to demonstrate that the AutoFi can support new types of gestures after self-supervised learning. The criterion is the top-1 accuracy across all test samples.

Implementation Details: Here, we introduce the details of the AutoFi, and the experimental settings are introduced in the following sections. The two modules of the AutoFi are implemented by Pytorch. The network structures are shown in Table I. The SGD optimizer is utilized with a learning rate of 0.01 and a momentum of 0.9. The epoch of training GSS module is 300 and the FSC is trained for 100 epochs. The batch size is set to 128 in order that the GSS module can capture the geometry among samples. For all the experiments, we set the hyper-parameter $\lambda = 1$ and $\gamma = 1000$, which aims to keep the magnitudes of multiple losses similar.

Baselines: As our method mainly deals with the few-shot learning scenario, we compare our method with recent state-of-the-art few-shot recognition methods based on CSI, including the CSI-GDAM [53], the ReWiS [54], and the classic prototypical network [50] that is the baseline method. The CSI-GRAM utilizes the graph neural network and attention scheme to enhance few-shot learning, while the ReWiS proposes SVD data processing and applies the prototypical network.

B. Real-World System Evaluation

System Setup: To demonstrate the effectiveness of our AutoFi, we implement our system in the real world. The AutoFi system consists of two TPLink-N750 routers that serve as the transmitter and receiver. They are set to operate on 5 GHz with a bandwidth of 40 MHz. Leveraging Atheros CSI tool [21] and real-time IoT platform [3], we extract the 114 subcarriers of CSI data for each pair of antennas. The receiver is equipped with three antennas while the transmitter is equipped with one antenna. The sampling rate is 100 Hz and each CSI sample is captured for 5 s with the size of $3 \times 114 \times 500$. Only CSI amplitudes are used since the phase information is not stable for the Atheros tool. As shown in Fig. 3, we evaluate the AutoFi in two different environments. The first environment only has one table and all chairs are surrounded, while the second one has a more complicated layout with four tables and many chairs. We set a threshold $\tau = 20$ to capture CSI samples randomly. As long as the CSI amplitude is greater than $\tau$, the system starts to record the CSI data for 5 s. In this way, we leave the AutoFi system alone for automatic data collection, and we obtain more than 5000 samples without any human labor for self-supervised learning. This automatic data collection process took about half a day. Then, we collect very few labeled CSI samples to conduct FSC, which can be easily achieved up to several minutes in the real world as only 1–3 samples are required for one gesture. The downstream tasks are the gesture recognition in the first environment and the human gait recognition in the second environment. The test samples are collected anywhere within the regions and they are annotated only for to serve as ground truth for performance. For gesture recognition, there are eight types of gestures including up & down, left & right, pull & push, clap, fist, circling, throw, and zoom, with 120 samples from each category for testing. For human identification, 14 volunteers are engaged with 20 samples from each category for testing. The volunteer walks through the line of sight of the two routers either with a jacket or a backpack, which makes the task challenging. Two experiments are independently conducted, and there exist some environmental dynamics as some staff are working around. No data preprocessing techniques are utilized for model training.

Results: According to different shots of FSC, we summarize the overall results in Table II. It is seen that the AutoFi achieves 83.31%, 87.46%, and 89.71% accuracy on gesture recognition task with 1-shot, 2-shots, and 3-shots learning, respectively, outperforming the baseline method by 4%–6%. For the human identification task, more categories and the heterogeneity of gaits lead to more challenges. The overall accuracy is worse than the accuracy in the gesture recognition task. The AutoFi still achieves state-of-the-art performance when it is compared to the ReWiS and CSI-GDAM. It is seen that the ReWiS only slightly outperforms the prototypical network, while the CSI-GDAM attains a stable improvement. In summary, the GSS module of the AutoFi learns the environmental dependency, and thus promotes the subsequent few-shot learning by the prototypical network. The results demonstrate that the AutoFi can learn randomly-collected samples by itself, and transfer the knowledge to distinct downstream tasks.

Feature Transferability: For human identification, we have three testing scenarios: 1) subjects wearing jacket; 2) subjects...
wearing backpacks; and 3) subjects wearing jacket and backpacks with enhanced environmental dynamics. We let the AutoFi only incorporate few-shot samples from a single scenarios and test it on all scenarios, which verifies the transferability ability of the features. We compare it with the single prototypical network in Table III. It is noted that our proposed AutoFi achieves significant improvements across all tasks. Especially, it improves the baseline method on one-shot learning for subjects in jacket by 27.25%. This demonstrates that the features learned by our method have strong transferability. Moreover, it is obvious that the situation of subjects in jacket has the best results for AutoFi. The reason is that the jacket or backpacks are interference in supervised learning, which may dominate the classifier. The learning-based models are prone to learn these irrelevant features because these may help identification but only for samples, not identity. For example, the backpack may swing as the subject passes by, which helps classification but not human identification. This further shows the importance of feature transferability and the negative effect of corrupt samples for normal few-shot learning.

C. Evaluation on UT-HAR Data Set

Data Setup: The UT-HAR [52] is a human activity recognition data set collected by the University of Toronto. There are seven categories, including lie down, fall, walk, run, sit down, stand up, and empty. The sampling rate is 1000 Hz that is too large for an input, and the data set is continuously without segmentation. Therefore, we can simulate our scenario by randomly segmenting the data set into pieces of CSI for the self-supervised training, and then conduct the few-shot testing. To this end, we segment the data randomly and get 3977 CSI samples. We prepare 10 and 20 labeled samples per category for FSC, and 70 samples per category for evaluation, which forms the 10-shots and 20-shots activity recognition problem. The size of the input data is 3 × 30 × 250. The first layer of the GSS module is slightly modified to match the input size.

Results: The results are shown in Fig. 4. The proposed AutoFi achieves the accuracy of 66.8% and 78.8% on 10-shots and 20-shots tasks, which demonstrates the effectiveness of our method. Nevertheless, the overall performances are lower than those of the real-world evaluation. The reason is twofold. First, the UT-HAR data set is a not well-segmented data set, so there still exists noises for few-shot training samples. Such noise hinders the training significantly. Second, the data set is collected using the Intel 5300 NIC [20] that only supports 30 subcarriers for each pair of antenna. The resolution is much lower than ours (i.e., 114 subcarriers). It is seen that the low resolution and data noises decrease the performance of few-shot learning.

D. Evaluation on Widar Data Set

Data Setup: Since the UT-HAR has intrinsic noises, we further investigate a large-scale data set collected by Intel 5300 NIC, the Widar [16]. In this data set, we directly use its transformed data, namely, body-coordinate velocity profile (BVP), which eliminates the influence of environment noises. The size of the BVP is 20 × 20 × T, and T = 40 is the duration. In this experiment, we aim to further demonstrate that the AutoFi helps increase the feature transferability in terms of new categories for other data modalities of CSI data. To this end, we use 16 categories of gestures for self-supervised learning and six gestures for FSC. The first layer of the GSS module is slightly modified to match the input size.
Fig. 5. Accuracy (%) comparison on Widar [16] data set.

Results: As shown in Fig. 5, the proposed AutoFi achieves 55.60% and 63.80% accuracy for 10-shots and 20-shots recognition tasks, respectively, outperforming the baseline method by 14.40% and 8.5%, respectively. It is observed that the ReWiS does not achieve improvement, and the possible reason is that the SVD method may not work for BVP. The overall performance on Widar is worse than that of UT-HAR and our real-world experiments since the testing data here does not come from one environment, which actually does not conform with our scenario. Nevertheless, we use this data set to demonstrate that the AutoFi can realize the enlargement of the gestures for the CSI-based gesture recognition system. Even though the training categories for the GSS are not overlapped with the testing categories and the environment varies, the AutoFi can still bring significant improvement for existing methods.

E. Ablation Study
To demonstrate the effectiveness of multiple objectives in the GSS module, we compare our method with the cases of the lack of mutual information loss and geometric consistency. The baseline performance has been illustrated in Table II, i.e., the prototypical network. Based on the real-world human identification experiments, we draw the results in Fig. 6. The “w.o.” denotes “without.” When the mutual information loss is absent, we can observe obvious performance decreasing for 2-shots and 3-shots cases. For 1-shot case, the performances are quite similar, because the scale of the training samples is rather limited. As for the geometric consistency, it leads to a marginal improvement for all scenarios, verifying its advantages for few-shot learning.

F. Model Inference and Time Cost
Though the AutoFi can learn CSI patterns by itself, model learning still requires computational resources. Here, we compute the training time and the model inference time for each CSI sample. In large-scale Wi-Fi sensing, these data can be easily uploaded and processed at the cloud, so we run all the program on a single NVIDIA RTX 2080Ti. For our real-time system, the GSS module learns 5000 CSI samples for 300 epochs, which cost 22 mins. The FSC module only takes less than 1 min. As this process is conducted offline, it is acceptable in reality. Compared to model training, we pay more attention to model inference for real-time systems. Our recognition model only costs 22 ms for one CSI sample in our system. For UT-HAR and Widar, as the data dimensions are lower, the cost time is only 16 and 15 ms, respectively. In this manner, we prove that the AutoFi can be easily setup and run efficiently in the real world.

V. Conclusion
In this article, we propose AutoFi, a novel GSS learning framework, which is the first work that realizes self-driven initialization of learning-based models using randomly-collected CSI data. The GSS learning enables the AutoFi to learn CSI patterns by consistency and mutual information, and an FSC module can efficiently empower the AutoFi to conduct downstream recognition tasks. Extensive experiments are conducted in both the real world and public data sets. The experimental results show that the AutoFi can significantly improve the few-shot performance, or enhance the existing systems by cross-task knowledge transfer. We believe that the AutoFi is an important step toward automatic and pervasive Wi-Fi sensing. Future works may focus on how to leverage limited labeled samples by exploiting data augmentation and how to integrate Wi-Fi and other modalities for robust sensing [55], [56].

REFERENCES
[1] D. J. Love, R. W. Heath, and T. Strohmer, “Grassmannian beamforming for multiple-input multiple-output wireless systems,” IEEE Trans. Inf. Theory, vol. 49, no. 10, pp. 2735–2747, Oct. 2003.
[2] Z. Yang, Z. Zhou, and Y. Liu, “From RSSI to CSI: Indoor localization via channel response,” ACM Comput. Surveys, vol. 46, no. 2, pp. 1–32, 2013.
[3] J. Yang, H. Zou, H. Jiang, and L. Xie, “Device-free occupant activity sensing using WiFi-enabled IoT devices for smart homes,” IEEE Internet Things J., vol. 5, no. 5, pp. 3991–4002, Oct. 2018.
[4] H. Wang et al., “Human respiration detection with commodity WiFi devices: Do user location and body orientation matter?” in Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., 2016, pp. 25–36.
[5] H. Zou, Y. Zhou, J. Yang, H. Jiang, L. Xie, and C. J. Spanos, “Deepsense: Device-free human activity recognition via autoencoder long-term recurrent convolutional network,” in Proc. IEEE Int. Conf. Commun. (ICC), 2018, pp. 1–6.
[54] N. Bahadori, J. Ashdown, and F. Restuccia, “ReWiS: Reliable Wi-Fi sensing through few-shot multi-antenna multi-receiver CSI learning,” 2022, arXiv:2201.00869.

[55] H. Zou, J. Yang, H. P. Das, H. Liu, Y. Zhou, and C. J. Spanos, “WiFi and vision multimodal learning for accurate and robust device-free human activity recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2019, pp. 426–433.

[56] L. Deng, J. Yang, S. Yuan, H. Zou, C. X. Lu, and L. Xie, “GaitFi: Robust device-free human identification via WiFi and vision multimodal learning,” 2022, arXiv:2208.14326.

Jianfei Yang (Member, IEEE) received the B.Eng. degree from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, in 2016, and the Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2021.

He used to work as a Senior Research Engineer with BEARS, University of California at Berkeley, Berkeley, CA, USA. His research interests include deep transfer learning with applications in Internet of Things and computer vision.

Dr. Yang won many AI and data challenges in the visual and interdisciplinary fields, such as ACM ICMI EmotiW-18, IEEE CVPR-19 UG2+ challenge, and ICCV-21 Masked Face Recognition challenge. He is currently an Independent Principal Investigator and a Presidential Postdoctoral Research Fellow at NTU.

Han Zou received the B.Eng. (First Class Hons.) and Ph.D. degrees in electrical and electronic engineering from the Nanyang Technological University, Singapore, in 2012 and 2016, respectively.

He is currently a Postdoctoral Scholar with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA, USA. His research interests include ubiquitous computing, statistical learning, signal processing and data analytics with applications in occupancy sensing, indoor localization, smart buildings, and Internet of Things.

Dazhuo Wang received the B.Eng. from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2018, where he is currently pursuing the Ph.D. degree with the School of Electrical and Electronic Engineering.

He is a Scholar of the Agency for Science, Technology and Research, Singapore, under AGS scholarship. His research interests include Industrial Internet of Things and machine learning.

Libhua Xie (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Newcastle, Callaghan, NSW, Australia, in 1992.

Since 1992, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he is currently a Professor and the Director of the Center for Advanced Robotics Technology Innovation. He served as the Head of Division of Control and Instrumentation and the Co-Director of Delta-NTU Corporate Laboratory for Cyber-Physical Systems. He held teaching appointments in the Department of Automatic Control, Nanjing University of Science and Technology, Nanjing, China, from 1986 to 1989. His research interests include robust control and estimation, networked control systems, multiagent networks, smart sensing, and unmanned systems.

Dr. Xie is the Editor-in-Chief for Unmanned Systems and has served as the Editor of IET Book Series in Control and an Associate Editor for a number of journals, including IEEE Transactions on Automatic Control, Automatica, IEEE Transactions on Control Systems Technology, IEEE Transactions on Network Control Systems, and IEEE Transactions on Circuits and Systems—II: Express Briefs. He was an IEEE Distinguished Lecturer from January 2012 to December 2014. He is a Fellow of Academy of Engineering Singapore, IEEE, IFAC, and CAA.

Xinyan Chen is currently an undergraduate student with the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore.

He worked for his Undergraduate Research Experience on Campus (URECA) Program under the supervision of Prof. Libhua Xie and Dr. Jianfei Yang at NTU. His research interests include deep learning and computer vision.

Libhua Xie received the Ph.D. degree in electrical engineering from the University of Newcastle, Callaghan, NSW, Australia, in 1992.

Since 1992, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he is currently a Professor and the Director of the Center for Advanced Robotics Technology Innovation. He served as the Head of Division of Control and Instrumentation and the Co-Director of Delta-NTU Corporate Laboratory for Cyber-Physical Systems.

He held teaching appointments in the Department of Automatic Control, Nanjing University of Science and Technology, Nanjing, China, from 1986 to 1989. His research interests include robust control and estimation, networked control systems, multiagent networks, smart sensing, and unmanned systems.

Dr. Xie is the Editor-in-Chief for Unmanned Systems and has served as the Editor of IET Book Series in Control and an Associate Editor for a number of journals, including IEEE Transactions on Automatic Control, Automatica, IEEE Transactions on Control Systems Technology, IEEE Transactions on Network Control Systems, and IEEE Transactions on Circuits and Systems—II: Express Briefs. He was an IEEE Distinguished Lecturer from January 2012 to December 2014. He is a Fellow of Academy of Engineering Singapore, IEEE, IFAC, and CAA.