SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

July 2026, Vol. 69, Iss. 7, 172303:1–172303:13
https://doi.org/10.1007/s11432-025-4745-6

Towards ubiquitous fingerprint-based localization with limited human effort

Lingyan ZHANG[1*], Taotao KONG[1], Yuanfeng QIU[3], Danyang QIN[1], Shaohua WU[2], Tingting ZHANG[2] & Qinyu ZHANG[2]

[1]*School of Electronics Engineering, Heilongjiang University, Harbin 150080, China*
[2]*School of Information Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China*
[3]*Information Hub, Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511400, China*

Received 22 May 2025/Revised 11 September 2025/Accepted 30 December 2025/Published online 8 June 2026

Abstract Recent fingerprint-based localization in a deep-learning framework has promising potential for providing location-based services with 6G's inherent intelligence. Classical deep neural networks are designed to extract features that can effectively improve accuracy in specific scenarios. However, the hurdles of radio annotation, fingerprint degradation, and feature dependence severely limit its universal applicability with unpredictable environmental dynamics. To address these issues, we propose a novel deep-learning-based localization framework to achieve ubiquitous location estimation with minimal human effort. The latent view-invariant factor underlying channel state information (CSI) data is refined via amplitude and phase sampling to establish a dual-view contrastive pretraining model, DVCLoc, that learns generic representations directly transferable to new environments or scenarios. Then, limited CSI fingerprints are used to train the location predictor for robust localization. Extensive real-world experiments demonstrate that DVCLoc achieves state-of-the-art localization performance across various complex scenarios, advancing DNN-based localization from specificity to generality.

Keywords location-based services, fingerprint-based localization, contrastive learning, CSI

Citation Zhang L Y, Kong T T, Qiu Y F, et al. Towards ubiquitous fingerprint-based localization with limited human effort. Sci China Inf Sci, 2026, 69(7): 172303, https://doi.org/10.1007/s11432-025-4745-6

1 Introduction

Because 6G requires inherently intelligent capabilities, fingerprint-based localization has become one of the most promising technologies for providing location-based services (LBSs) and location-aware applications [1–5]. It usually collects extensive radio fingerprints to construct a radio map and extract the features that can effectively associate radio signals with their corresponding location information. Then, the user's location is estimated by best-fitted fingerprinting [5]. Compared with other localization approaches based on sophisticated signal estimations, such as the time of flight (ToF) [6], angle of arrival (AoA) [7], or hybrid measurements [8], the WiFi-based localization solution has the advantages of easy signal availability, simple algorithm implementation, and satisfactory localization accuracy. These advantages make this solution appealing to become the first large-scale commercial deployment in various scenarios [9, 10], such as Baidu Map, Google Map, and OpenStreetMap.

Mainstream localization approaches prefer classical deep neural networks (DNNs) [11–13] to learn implicit representations that can effectively improve accuracy in certain scenarios. DeepFi [11] has designed the encoder-decoder to learn unsupervised features and determine the location estimation in a probabilistic method. WiDeep [12] employs a stacked denoising auto-encoder to learn noise-resilient representations with complex environmental changes. CNNLoc [13] adopts conventional neural networks (CNNs) to learn spatial-related representations for accurate location estimations. However, the following drawbacks hinder these DNN-based solutions from providing intelligent LBSs (ILBSs) for future networks. First, it is difficult to annotate radio signals with accurate location information, since they are not human-interpretable [14]. In a data-driven manner, annotating massive radio signals requires immeasurable human effort to construct a radio map with an extensive site survey. A crowdsourcing-based technique has been adopted to automatically collect radio fingerprints via multi-sensor data fusion, leading to noisy location annotations with the heterogeneous measurements [15, 16]. Second, the radio features learned by vanilla DNNs are environment-dependent and lack environmental adaptability, leading to inaccurate localization. Since received

* Corresponding author (email: lingyan.z@hlju.edu.cn)

signal strengths (RSSs) are extremely vulnerable to environmental dynamics, the fingerprint degradation over time incurs a mismatch between current fingerprints and the existing radio map, and even results in location estimation failures [16–18], especially in indoor localization. The intuitive solution is to periodically update the radio map through continuous fingerprint collection and calibration, which are also costly in terms of system maintenance. Last but not least, it is a key challenge to improve the generalization capability for ubiquitous ILBSs, like the global positioning system (GPS) in outdoor scenarios [19]. In a new scenario, different spatial layouts induce distinct wireless propagation, which requires the rebuilding of a radio map to retrain the model from scratch, severely restricting the universal applicability of deep-learning-based localization systems [20].

In this paper, we propose a novel deep learning-based localization framework to achieve ubiquitous location estimation with minimal human effort. The latent invariant factor underlying radio data is explored and exploited to learn the generic-purpose representations that can be effectively generalized to new environments or new scenarios. We fully utilize the physical layer channel state information (CSI) via the amplitude and phase measurements, which enable the same state of the target to be recorded with the intrinsic dual-view sampling. Although each view's sampling is noisy due to hardware imperfections and unpredictable environmental dynamics [21, 22], the crucial information tends to be shared between the views. On this basis, we develop the dual-view contrastive pretraining model, DVCLoc, to learn generic representations by maximizing feature similarities in a low-dimensional space. For downstream localization tasks, the learned feature encoders are directly transferred, and the location predictor is trained using limited CSI fingerprints to achieve robust localization. Extensive experiments evaluation show that DVCLoc achieves state-of-the-art (SoTA) localization performance across various complex scenarios.

Our main contribution can be summarized as follows.

(1) We propose a novel framework of fingerprint-based localization to establish a dual-view contrastive pretraining model for learning the generic representations of radio data, and further, DVCLoc achieves GPS-like localization with limited human effort.

(2) We provide a fresh perspective on learning feature representations by refining the latent invariant factor underlying the radio samplings of different views. This approach can stay robust in the face of complex environmental dynamics, thereby addressing the limitations of environmental adaptability and system generalization.

(3) We validate the effectiveness of the DVCLoc prototype in typical indoor scenarios with extensive performance evaluations. DVCLoc achieves SoTA performance with the trade-offs of localization accuracy, cost-effective implementation, and generalization, making great progress in deep-learning-based localization technology from specificity to generality.

2 Related work

In order to provide practical LBSs, fingerprint-based localization schemes have recently endeavored to acquire environmental adaptability by refining a robust factor underlying radio fingerprints, which can be roughly divided into two categories: fingerprint-level and algorithm-level.

In the fingerprint level, the robust factor can be obtained by comparing RSSs with the neighboring locations. While RSS fluctuations, deteriorated with multipath propagation, are severe with complex environmental changes [16], the local proximity of the neighboring fingerprints remains relatively stable. Differential fingerprint patterns [16] and RSS gradient fingerprints [23, 24] are proposed to realize robust localization without any additional auxiliary. Furthermore, GraphLoc [25] designs multilayer graph attention networks with the residual structure to encode the structural information of radio fingerprints for adaptive localization. In order to achieve reasonable localization accuracy, however, these solutions still require an extensive site survey to perform an initial system deployment with the time-consuming effort.

In terms of the algorithm level, the advanced DNN framework is harnessed to learn environment-invariant representations with the robustness of environmental dynamics [4, 17, 26–28]. CSI-SCNN [26] and LESS [17] design Siamese neural networks to learn the similarity metric of feature embeddings, which model the neighboring relationships among different radio fingerprints. MetaLoc [20] adopts model-agnostic meta-learning to learn generalizable features with a small amount of radio fingerprints. Alternatively, transferable-knowledge-assisted localization schemes [18, 27, 28] attempt to learn transferable representations through the minimization of the distribution discrepancy before and after environmental dynamics. RobLoc [27] quantifies the feature distribution discrepancy with the empirical maximum mean discrepancy (MMD) distance and further incorporates it into DNN for automatic radio map adaptation. DadLoc [28] develops dynamic adversarial adaptation networks to learn the evolving representations by a mini-max game. While the finer representations are learned to attain performance enhancement, recent advances have not only required an initial radio map but also developed the complex network architecture,

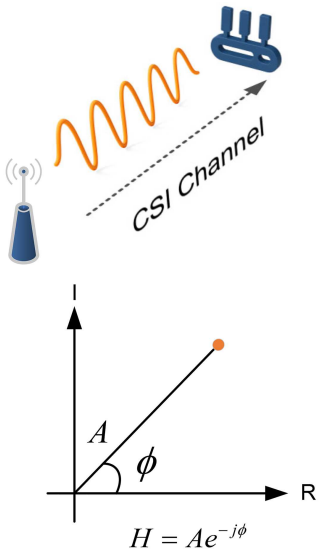

Figure 1 (Color online) CSI with intrinsic dual-view samplings.

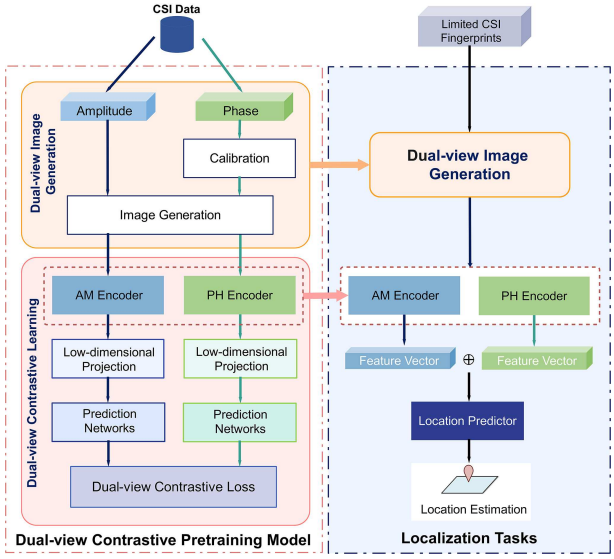

Figure 2 (Color online) DVCLoc's overview.

leading to the increasing computational overhead.

Above all, we propose a novel perspective of learning feature representations by mining the latent invariant factor underlying radio data via dual-view sampling, effectively overcome poor generalization with limited human efforts.

3 Problem formulation and motivation

WiFi fingerprint-based localization approaches are usually based on the location information supervision in two stages: offline training and online localization. During the offline training, radio fingerprints are collected at the pre-deployed reference points (RPs) to construct a radio map, i.e., $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{K}$ where $\mathbf{x}_i$ is the radio vectors with the corresponding location label y_i in the interested environment. In a deep learning framework, classical DNNs [11–13] are designed to extract the appropriate features $f(\mathbf{x})$ with the learnable parameters θ. Then, the location predictor $\hat{y} = p\left(f(\mathbf{x})\right)$ is trained with the softmax classification. The DNN-based localization model is formulated as

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{K} \sum_{i=1}^{K} \ell_{ce}\left(p(f(\mathbf{x}_i)), y_i\right), \tag{1}$$

where ℓ_{ce} is the cross-entropy loss. For accurate localization, it is inevitable to collect a great number of radio fingerprints, which requires time-consuming and labor-intensive human effort [5], worse for the DNN-based localization in a data-driven manner [4]. Moreover, DNN-based approaches have learned the environment-dependent representations without the robustness of environmental dynamics [17], severely restricting their universal applicability.

To address these issues, we mine the invariant factor underlying radio data to learn generic-purpose representations for the effectiveness of improving system generalization. As illustrated in Figure 1, we resort to CSI readings including both amplitude and phase measurements, which can record the same state of the target with the intrinsic dual-view sampling. With WiFi MIMO-OFDM communication, the CSI $\mathbf{H} \in \mathbb{C}^{M\times N}$ is received with the complex values as

$$\mathbf{H} = \mathbf{A}\mathrm{e}^{-\mathrm{j}\boldsymbol{\Phi}}, \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{\Phi} \in \mathbb{R}^{M\times N}$ are the amplitude and phase matrices from the N subcarriers over the M antennas, respectively. Although each view's sampling is noisy due to hardware imperfections [21, 22], the crucial information tends to be shared between the views with unpredictable environmental dynamics. Therefore, we propose the DVCLoc's pretraining model to learn the view-invariant representations that encode the co-occurrence information of dual-view data, essentially achieving the effectiveness of generalizable capability for ubiquitous localization across diverse complex scenarios.

4 System design and implementation

As illustrated in Figure 2, the proposed DVCLoc system first establishes the dual-view contrastive pretraining model to learn generic representations underlying CSI dual-view data and then accomplishes the localization tasks across various complex scenarios for offering ubiquitous ILBSs.

The first stage is to establish the dual-view contrastive pretraining model for general-purpose feature learning. In data preprocessing, since raw CSI measurements are extremely noisy with uncertain interferences, we calibrate them to obtain the clean dual-view samples. Then, the dual-view image generation is performed to improve the quality of learning feature representations. Then, we propose the framework of the dual-view contrastive learning to learn the view-invariant representations without the location information supervision. DVCLoc develops Siamese convolutional networks to train the dual-view feature encoders, such as the AM encoder for the amplitude samples and the PH encoder for the calibrated phase samples, respectively. The low-dimensional projection and the prediction networks are designed to enable the two-branch networks to interact and learn from each other. In the latent representation space, the dual-view contrastive loss is formulated to maximize the similarities between the dual-view samples.

For localization tasks, the learned dual-view feature encoders are frozen to directly obtain the feature vectors. Then, limited CSI fingerprints are used to train the location predictor at a lower deployment cost. When a real-time CSI query is received online to generate the dual-view images, the dual-view features are captured and combined to determine the location estimation in a probabilistic method.

4.1 Dual-view contrastive pretraining model

4.1.1 *Dual-view image generation*

By using off-the-shelf WiFi infrastructures [22, 29], CSI measurements contain the amplitude and phase measurements of multi-antenna and multiple sub-carriers, which naturally own the double views to record the same state of the target. However, raw CSI samples of each view are noisy with uncertain interferences. We need to calibrate them and then utilize the temporal-spatial correlation of CSI readings to generate the dual-view images, which can improve the quality of the learned representations in our dual-view contrastive pretraining model.

Calibration. Due to hardware imperfections between the transceivers, raw phase responses are severely noisy with random phase offsets [21, 22]. The sampling time offset (STO) brings a constant offset to every CSI packet in an array, which leads to a linear trend of the additional phase shifts in the frequency terms of the multiple subcarriers [21, 29]. We use the optimal linear fitting to remove the STO for CSI calibration.

The raw phase measurements $\hat{\phi}(m,n)$ are received at the nth subcarrier in the mth antenna as

$$\hat{\phi}(m,n) = \phi(m,n) + 2\pi f_n \tau_s + z, \tag{3}$$

where $\phi(m,n)$ is the true phase response caused by radio propagation in the interested environment, f_n is the frequency of the nth subcarrier, τ_s represents the delay caused by the STO, and z is defined as the measurement noise. When the radio signals are propagated with environmental dynamics, the random phase offset at the nth subcarrier is caused as $2\pi(n-1)f_\delta\tau_s$ where f_δ denotes the frequency interval between the adjacent subcarriers. Since the random phase offset in an array is the same for every antenna, we adopt the optimal linear fitting [29] to eliminate random phase offsets. The linear fitting parameters $\hat{\tau_s}$ and $\hat{\varepsilon}$ are worked out as

$$(\hat{\tau_s}, \hat{\varepsilon}) = \arg\min_{\tau_s,\varepsilon} \sum_{m,n=1}^{M,N} \left(\hat{\phi}(m,n) + 2\pi(n-1)f_\delta\tau_s + \varepsilon\right)^2, \tag{4}$$

where $\hat{\phi}(m,n)$ is the unwrapped phase at the nth subcarrier in the mth antenna. We can turn away the phase offsets to capture the calibrated phase responses $\bar{\phi}(m,n)$ as

$$\bar{\phi}(m,n) = \hat{\phi}(m,n) - 2\pi(n-1)f_\delta\hat{\tau}_s - \hat{\varepsilon}. \tag{5}$$

Image generation. In order to learn finer representations, we utilize the temporospatial correlation of CSI reading with MIMO-OFDM to transform the CSI amplitudes and the calibrated phases into radio images [18, 28]. Due to the different scales, we apply the min-max normalization to both the CSI amplitudes and the calibrated phases. Given the amplitude vector $A = [A_1, A_2, \ldots, A_N]$ at each antenna, the amplitude value A_n at the nth subcarrier is normalized as

$$\tilde{A}_n = \frac{A_n - A_{\min}}{A_{\max} - A_{\min}}, \tag{6}$$

where $A_{\min}$ and $A_{\max}$ are the minimal and maximal values of the amplitude vector A in the same antenna. Meanwhile, the calibrated phase value $\bar{\phi}_n$ at the nth subcarrier is normalized as

$$\tilde{\phi}_n = \frac{\bar{\phi}_n - \bar{\phi}_{\min}}{\bar{\phi}_{\max} - \bar{\phi}_{\min}}, \tag{7}$$

where $\bar{\phi}_{\min}$ and $\bar{\phi}_{\max}$ are the minimal and maximal values of the calibrated phase vector $\bar{\Phi} = [\bar{\phi}_1, \bar{\phi}_2, \ldots, \bar{\phi}_N]$ in the same antenna.

Furthermore, the dual-view dataset is constructed with the normalized CSI data. We transform the normalized CSI amplitudes at each antenna into an individual radio image and stitch these images from one array to create the amplitude images, similarly with the generation of the corresponding phase images.

4.1.2 *Dual-view contrastive learning*

In our dual-view contrastive pretraining model, we feed the unannotated dual-view images into the developed Siamese convolutional networks to learn the view-invariant representations through their similarity maximization in the latent space. The details are elaborated as follows.

Feature encoder. CNNs are designed as the feature encoder $f(\mathbf{x})$

$$\mathbf{v} = f(\mathbf{x}) = \mathrm{CNN}(\mathbf{x}), \tag{8}$$

where $\mathbf{x}$ is the input images. According to the great capability of CNN's image representations, the feature vector $\mathbf{v}$ is trained by the CNN block, which has two convolutional layers and one max-pooling layer. We develop the twin-branch networks to perform dual-view contrastive learning, and thus, one feature encoder trains the feature vectors of the amplitude images, as the AM encoder, and the other is the PH encoder, correspondingly. Once pretrained with network convergence, the double feature encoders are transferable to the downstream tasks of robust localization.

Low-dimensional projection. The feature vectors are further projected into a low-dimensional space by the nonlinear transformation, facilitating the two-branch networks to interact and learn from each other. We leverage the multilayer perceptron (MLP) with two fully connected (FC) layers to obtain the feature embeddings as

$$\mathbf{h} = h(\mathbf{v}) = \mathrm{MLP}(f(\mathbf{x})). \tag{9}$$

The amplitude feature embeddings $\mathbf{h}^a$ and the phase feature embeddings $\mathbf{h}^p$ from the same CSI measurement are defined, respectively.

Dual-view contrastive loss. In the low-dimensional space, the dual-view feature embeddings from the same CSI sample should be as consistent as possible by maximizing their similarities. Under the cross-view prediction [30], the feature embeddings of one view should be matched with the other view. The negative cosine similarity is adopted to evaluate the similarity between the one view's embedding prediction $\mathbf{z}^1 = g(\mathbf{h}^1)$ and the other view's embedding $\mathbf{h}^2 = h\left(f(\mathbf{x}_2)\right)$ as

$$\mathrm{sim}(\mathbf{z}^1, \mathbf{h}^2) = -\frac{\mathbf{z}^1}{\|\mathbf{z}^1\|_2} \cdot \frac{\mathbf{h}^2}{\|\mathbf{h}^2\|_2}, \tag{10}$$

where $\|\cdot\|_2$ is ℓ_2-norm. The symmetrized similarity between the dual-view feature embeddings is formulated as

$$\mathcal{L}_i = \frac{1}{2}\mathrm{sim}(\mathbf{z}_i^a, \mathbf{h}_i^p) + \frac{1}{2}\mathrm{sim}(\mathbf{z}_i^p, \mathbf{h}_i^a). \tag{11}$$

To avoid collapsed representations, we use the stop-gradient operation sg of one branch network without the feature prediction [31]. The loss of dual-view contrastive learning is rewritten as

$$\mathcal{L}_i = \frac{1}{2}\mathrm{sim}(\mathbf{z}_i^a, \mathrm{sg}(\mathbf{h}_i^p)) + \frac{1}{2}\mathrm{sim}(\mathbf{z}_i^p, \mathrm{sg}(\mathbf{h}_i^a)). \tag{12}$$

Note that the AM encoder receives no gradient from the phase embeddings $\mathbf{h}^p$ on the amplitude feature prediction. It only receives the gradients of the amplitude's embedding from $\mathbf{z}^a$, and vice versa for the PH encoder. The overall loss of our dual-view contrastive pretraining model is averaged with all CSI training data $\mathcal{D}_{pretrain} = \{\mathbf{x}_i\}_{i=1}^K$ as

$$\mathcal{L} = \sum_{i=1}^{K} \left(\frac{1}{2}\mathrm{sim}(\mathbf{z}_i^a, \mathrm{sg}(\mathbf{h}_i^p)) + \frac{1}{2}\mathrm{sim}(\mathbf{z}_i^p, \mathrm{sg}(\mathbf{h}_i^a)) \right). \tag{13}$$

According to the similarity maximization, DVCLoc attempts to capture the view-invariant information underlying unannotated CSI data for learning generic representations, effectively enhancing the generalizable capability.

Training strategy. We clarify the DVCLoc's update with the similarity maximization between the feature embeddings of dual-view images. From one view, the feature embedding is obtained by the low-dimensional projection of the feature vectors as $\mathbf{h} = h_\theta(f_\theta(\mathbf{x}^1))$ and the prediction network outputs its feature prediction $\mathbf{z} = z_\theta(\mathbf{h})$. For view-invariant representation learning, the feature prediction of one view's embedding should be similar to the other view's features $\mathbf{h}' = h_\eta(f_\eta(\mathbf{x}^2))$ with the learnable parameter set η, which is formulated as

$$\mathcal{L}(\theta,\eta) \triangleq \mathbb{E}_{\mathbf{x^1},\mathbf{x^2}}\left\{\left\|\hat{z}_\theta\left(h_\theta\right)-\hat{h}'_\eta\right\|_2^2\right\}, \tag{14}$$

where $\hat{z}_\theta = \frac{z_\theta}{\|z_\theta\|_2}$ and $\hat{h}'_\eta = \frac{h_\eta}{\|h_\eta\|_2}$ are the normalization terms, and $\mathbf{x}^1 = \mathbf{x}^a$, $\mathbf{x}^2 = \mathbf{x}^p$. This loss function of the mean square error is equivalent to the negative cosine similarity as (10) [30, 31]. During each training iteration, the stochastic optimization is performed to minimize the loss through the low-dimensional network and the feature prediction, but without the gradient from the feature embeddings $\hat{h}'_\eta$. Similarly, the contrastive loss $\bar{\mathcal{L}}_{\theta,\eta}$ in the other branch is defined as

$$\bar{\mathcal{L}}(\theta,\eta) \triangleq \mathbb{E}_{\mathbf{x^1},\mathbf{x^2}}\left\{\left\|\hat{z}'_\eta\left(h_\eta\right)-\hat{h}_\theta\right\|_2^2\right\}. \tag{15}$$

For symmetrization, the overall loss of our pretraining model is formulated as

$$(\hat{\theta},\hat{\eta}) = \arg\min_{\theta,\eta}\mathcal{L}^{\mathrm{DVC}}_{\theta,\eta} = \arg\min_{\theta,\eta}\left(\mathcal{L}_{\theta,\eta}+\bar{\mathcal{L}}_{\theta,\eta}\right). \tag{16}$$

During our dual-view contrastive pretraining, we adopt an alternating optimization to learn view-invariant representations, which fixes one set of the learnable parameters and updates the other set. The AM encoder is first trained by the learnable parameters θ with stochastic gradient descent (SGD), but there is no gradient from the low-dimensional projection of the phase features $h_\eta\left(f_\eta(\mathbf{x}^p)\right)$. By following (14), we calculate the empirical expectation $h'_\eta = \mathbb{E}\left[h_\eta\left(f_\eta(\mathbf{x}^p)\right)\right]$ with the learnable parameters, and the parameter set of the AM encoder updates as

$$\theta \longleftarrow \theta - \mu\frac{\partial\mathcal{L}(\theta,\eta)}{\partial\theta}, \tag{17}$$

where μ is the learning rate. Secondly, the PH encoder is optimized with the learnable parameter set η by (15). Under the loss of $\bar{\mathcal{L}}(\theta,\eta)$, DVCLoc updates the corresponding parameters by the feature prediction $\hat{z}'_\eta(h_\eta)$ and the embedding expectation of the amplitude images $h_\theta = \mathbb{E}\left[h_\theta\left(f_\theta(\mathbf{x}^a)\right)\right]$ with the stop-gradient as

$$\eta \longleftarrow \eta - \mu\frac{\partial\bar{\mathcal{L}}(\theta,\eta)}{\partial\eta}. \tag{18}$$

4.2 Localization tasks

For accurate localization across diverse scenarios, the learned feature encoders are directly transferred for downstream tasks, and the location predictor is trained with limited CSI fingerprints at lower deployment effort. Finally, the location estimation is determined in a probabilistic method.

Given limited CSI fingerprints $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^k$, $k < K$, with the location labels y. The learned feature encoders are directly used to obtain the feature vectors of the dual-view samplings $\mathbf{v}^a$ and $\mathbf{v}^p$. We linearly combine the dual-view feature representations, as $\mathbf{v}_{\mathrm{com}} = f(\mathbf{x}^a) + f(\mathbf{x}^p) = \mathbf{v}^a + \mathbf{v}^p$, to feed into the location predictor. With the softmax classification, the training loss of the location predictor p is formulated as

$$\mathcal{L}_p = \sum_{i=1}^{k}\ell_{ce}\left(p\left(f(\mathbf{x}_i^a)+f(\mathbf{x}_i^p)\right), y_i\right), \tag{19}$$

where the output of the location predictor is $\hat{y} = p\left(f(\mathbf{v}_{\mathrm{com}})\right)$.

In the localization stage, receiving CSI measurements are transformed into the dual-view images in a specific scenario. The frozen feature encoders capture the dual-view feature vectors $\mathbf{v}^a$ and $\mathbf{v}^p$, respectively. Their combined features $\mathbf{v}_{\mathrm{com}}$ are obtained to predict the location labels. With the softmax regression, the posterior probability of the location predictor $\Pr(y|\mathbf{v}_{\mathrm{com}})$ is calculated to determine the location estimation $\hat{L}$ as

$$\hat{L} = \sum_{i=1}^{k}\Pr(y|\mathbf{v}_{\mathrm{com}})L_y, \tag{20}$$

where L_y is the location information with the corresponding location label y_i.

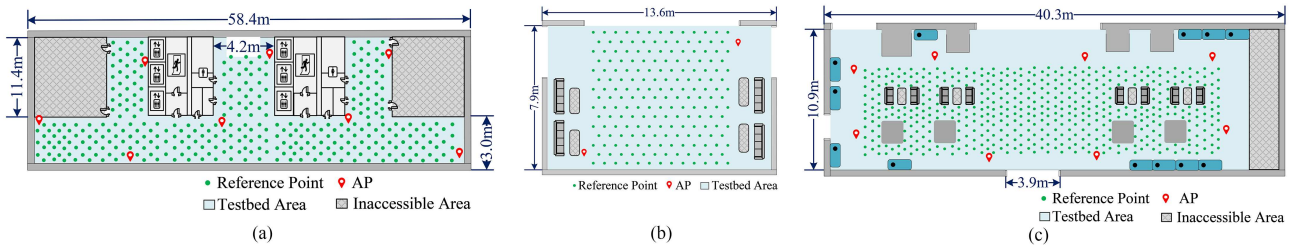

Figure 3 (Color online) Floor layouts of indoor scenarios. (a) Corridor; (b) hall; (c) lounge.

5 Experimental evaluation

In this section, we validate and evaluate the effectiveness of the proposed DVCLoc prototype with extensive real-world experiments. We introduce the experimental settings and methodology, and then discuss the DVCLoc system's performance across diverse indoor scenarios.

5.1 Experiment methodology

5.1.1 *CSI collection and indoor scenarios*

We use off-the-shelf WiFi infrastructure with commercial 802.11n 5300 NICs to collect CSI data in three typical indoor scenarios, as illustrated in Figure 3. The WiFi transceivers are set by the injection mode in the 5.32 GHz spectrum, and further, CSI readings are recorded by the receiver monitoring using the Linux CSI tool [32] in the Ubuntu 20.04 system. In the testbed scenarios, the predefined RPs are placed to collect CSI data. We also record all test CSI data with their corresponding location information to further evaluate localization performance in various indoor scenarios. The first scenario is the corridor in Figure 3(a), which is the size of 298 m^2 with dense multipath propagation. We randomly select 359 RPs to collect CSI data for the training dataset. There are 360 in the accessible area. The second scenario is the hall, which has a size of 107 m^2, with much line-of-sight (LOS) propagation. There are 188 RPs deployed as the training points and 187 testing points. In the last scenario, the lounge has a size of 436 m^2 where the furniture and obstacles cause complicated wireless propagation with the non-LOS (NLOS) path. We set 425 RPs and 426 test locations to evaluate the localization performance. We record CSI readings from every predefined RP, as CSI fingerprints. The CSI fingerprint database from all RPs is constructed via a site survey. We subsequently conduct such site surveys in all experimental areas with the environment changing. To evaluate robustness performance, the procedures for all CSI fingerprint collection have been repeated over a period of 8 months, during which radio signals fluctuate due to long-term environmental changes, such as temperature, humidity, and other weather changes. Environmental dynamics also occur with furniture changes, the user walking, and frequent switching doors and windows over time.

5.1.2 *Network architecture and pretraining strategy*

As shown in Figure 4, DVCLoc's architecture comprises dual-view contrastive learning and downstream localization tasks. In the proposed dual-view contrastive pretraining model, the backbone networks are trained to learn the generic representations using Algorithm 1. For 60 epochs, we maintain a batch size of 1024 with batch normalization. Both feature encoders are designed using a convolutional block, which consists of two convolutional layers, a ReLU nonlinearity, and a max-pooling layer. This convolutional block uses four 3×3 kernels with a stride of 1 and a padding of 1, and captures the 100-dimensional feature vectors from the generated dual-view images. Then, the dimensionality of the feature vectors is reduced via the low-dimensional projection using two FC layers with the sizes $\{100, 30\}$. Pretraining optimization is performed by using SGD with an initial learning rate of 0.05 and a weight decay of 5×10^{-4}. For downstream localization tasks, the location predictor is trained using a softmax layer with the appropriate number of RPs across various indoor scenarios.

We use all the CSI training datasets from three typical indoor scenarios to train our dual-view contrastive pretraining model. The two-branch networks are fed by generating the dual-view samplings from the same CSI data, which contain amplitude and phase measurements. There are 28000 CSI samples, and the model size is 32440. Network convergence costs 29.51 s of pretraining time to learn view-invariant representations for downstream localization tasks. Then, the learned feature encoders are directly transferred to downstream tasks. The corresponding location information of all test location points is recorded to evaluate their localization accuracy in every specific scenario.

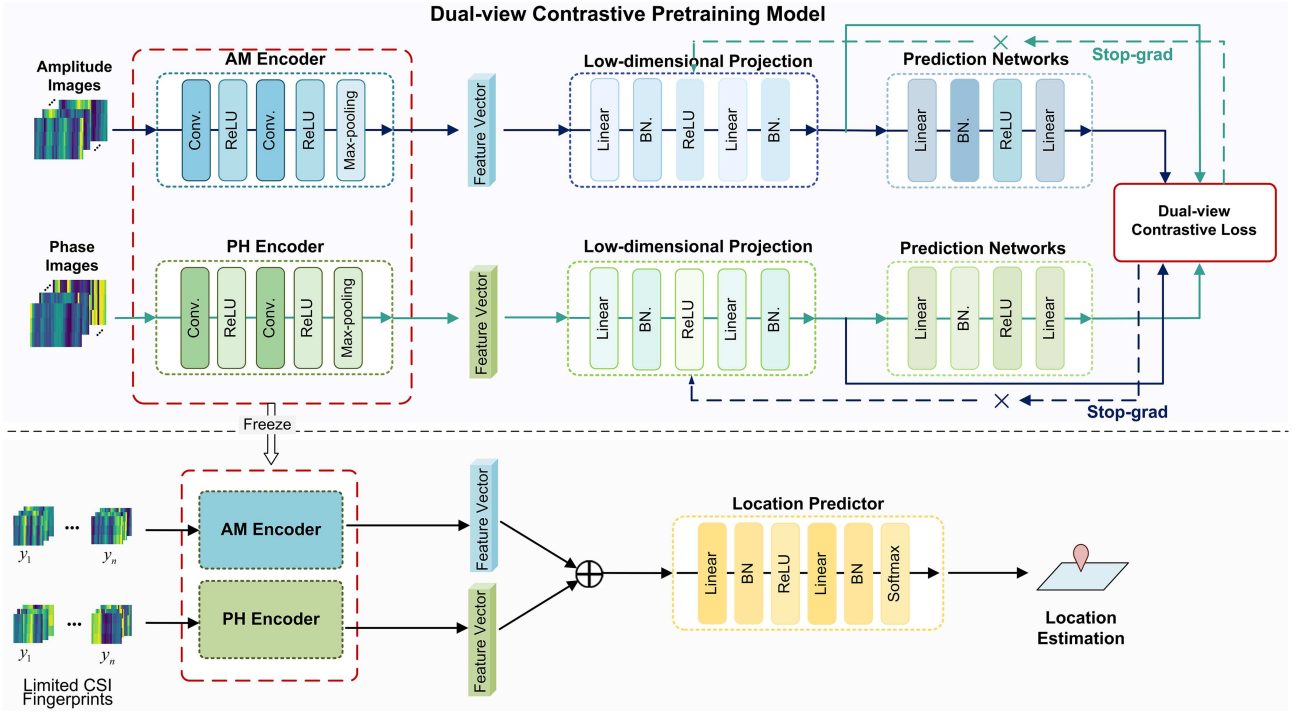

Figure 4 (Color online) DVCLoc's architecture.

Algorithm 1 Dual-view contrastive pretraining model.

Input: Unannotated CSI images $\mathcal{X} = \{\mathbf{x}_i^a, \mathbf{x}_i^p\}_{i=1}^K$ via the dual-view samplings, and the batch size B.
Output: The dual-view feature encoders f_θ and f_η for downstream localization tasks.

Initialize: The learnable parameters $\{\theta, \eta\}$ in the two branches, respectively;
for Each training iteration **do**
  Capture the feature vectors $\mathbf{v}^a = f_\theta(\mathbf{x}_i^a)$ and $\mathbf{v}^p = f_\eta(\mathbf{x}_i^p)$;
  Obtain the feature embeddings $\mathbf{h}^a = h_\theta(\mathbf{x}_i^a)$ and $\mathbf{h}^p = h_\eta(\mathbf{x}_i^p)$;
  Obtain the feature predictions $\mathbf{z}^a = z_\theta(\mathbf{h}^a)$ and $\mathbf{z}^p = z_\eta(\mathbf{h}^p)$;
  Calculate the symmetric loss with Eq. (12);
  Calculate the overall loss of the pretraining model as Eq. (13) with the batch size;
  Update θ and η by Eqs. (17) and (18);
end for
Return f_θ, f_η.

5.1.3 *Comparative methods*

To validate the effectiveness of our DVCLoc scheme, we compare its localization performance to classical and SoTA deep-learning-based schemes.

- Supervised learning-based methods: WiDeep [12], CNNLoc [13], GNN-Loc [33], and DeepSense [34].
- Unsupervised learning-based method: DeepFi [11].
- Semi-supervised learning-based method: ReNet-Loc with relation networks [17].

For fairness, we use the same convolutional structure to extract feature vectors across all localization solutions. As illustrated in Figure 4, the feature encoders of the proposed DVCLoc scheme comprise two convolutional layers and one max-pooling layer, which are also used in other comparative methods, such as CNNLoc, ReNet-Loc, and DeepSense. We also design two-layer graph attention networks for GNN-Loc and two fully-connected layers for DeepFi and WiDeep.

5.1.4 *Evaluation metric*

The following wide criteria of performance evaluation are used:

- Root mean squared error (RMSE) $\hat{\epsilon} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\hat{L}_i - L_i\right)^2}$;
- Mean absolute error (MAE) $\bar{\epsilon} = \frac{1}{k}\sum_{i=1}^{k}|\hat{L}_i - L_i|$;
- Standard deviation (STD) $\sigma = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(\epsilon_i - \bar{\epsilon})^2}$, where $\epsilon_i = |\hat{L}_i - L_i|$.

Table 1 Comparison of localization accuracy with different DNN-based approaches.

Method	RMSE $\hat{\epsilon}$ (m)	MAE $\bar{\epsilon}$ (m)	STD σ (m)
WiDeep	2.64	2.31	1.81
CNNLoc	2.50	2.32	1.75
GNN-Loc	2.35	2.10	1.64
DeepSense	2.65	2.25	1.39
DeepFi	2.72	2.32	1.84
ReNet-Loc	2.43	2.13	1.71
DVCLoc	2.13	1.79	1.15

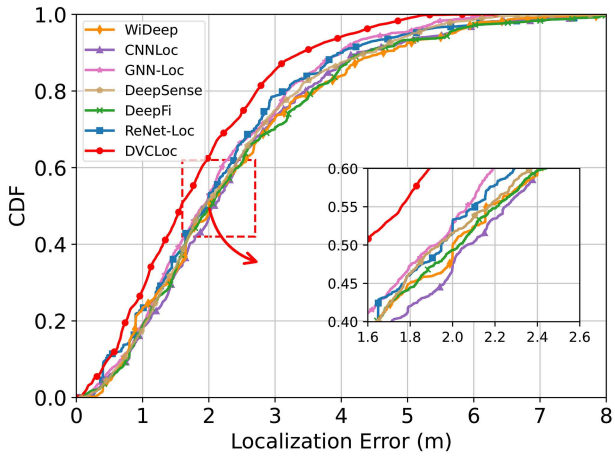

Figure 5 (Color online) Overall localization accuracy comparison.

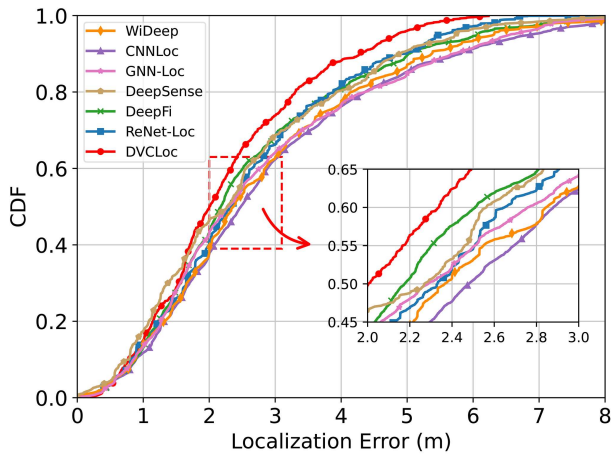

Figure 6 (Color online) Robust localization accuracy comparison.

5.2 Performance evaluation and discussion

5.2.1 *Localization accuracy*

We first calculate and compare the localization accuracy of the proposed DVCLoc prototype against several SoTA approaches in these typical indoor scenarios. As illustrated in Figure 5, the cumulative distribution functions (CDF) of localization errors are shown by the RMSEs of the proposed DVCLoc and other comparative schemes. DVCLoc achieves a median localization error of 1.58 m with accuracy improvements of 21% over WiDeep, 23.3% over CNNLoc, 17.7% over GNN-Loc, and 18.8% over ReNet-Loc. At the 80th percentile accuracy, DVCLoc outperforms WiDeep by 24.7%, CNNLoc by 21.3%, GNN-Loc by 14.4%, and ReNet-Loc by 13.4%, respectively. By the linear fusion with the amplitude and phase embeddings, the median localization error of DVCLoc is 18.1% lower than that of DeepSense. With these localization results, DVCLoc demonstrates the accuracy outperformance of the supervised learning-based approaches. Unlike supervised methods, DVCLoc leverages dual-view contrastive learning to refine the latent invariance underlying unlabeled CSI data, thereby effectively improving generalization capability for robust indoor localization across various complex scenarios. The pretraining model can capture the intrinsic data correlations that are often overlooked in supervised learning. Moreover, compared with deep unsupervised learning, DVCLoc achieves a median accuracy of 22% higher than DeepFi and an improvement of 23.5% at the 80th percentile.

Table 1 summarizes the localization performance compared with other evaluation metrics. With the overall RMSE evaluation, the proposed DVCLoc system reduces the location errors by 19.3% for WiDeep, 14.8% for CNNLoc, 9% for GNN-Loc, 12.3% for ReNet-Loc, 19.6% for DeepSense, and 21.6% for DeepFi, respectively. In addition, DVCLoc achieves a lower STD of 1.15 m, yielding more robust location estimations than others in diverse indoor scenarios. These results demonstrate that the proposed DVCLoc system can enhance DNN-based localization, achieving effective accuracy improvement for providing practical ILBSs.

5.2.2 *Robustness performance*

With the complex environmental dynamics, we collected CSI fingerprints at different time points over approximately 8 months to assess the robustness performance. As localization results shown in Figure 6, DVCLoc also demonstrates superior accuracy over other schemes with a median error of 2.0 m, achieving an accuracy improvement of 15.2% for WiDeep, 19% for CNNLoc, 12.2% for GNN-Loc, 11.5% for DeepSense, 7.8% for DeepFi, and 14.1% for ReNet-Loc.

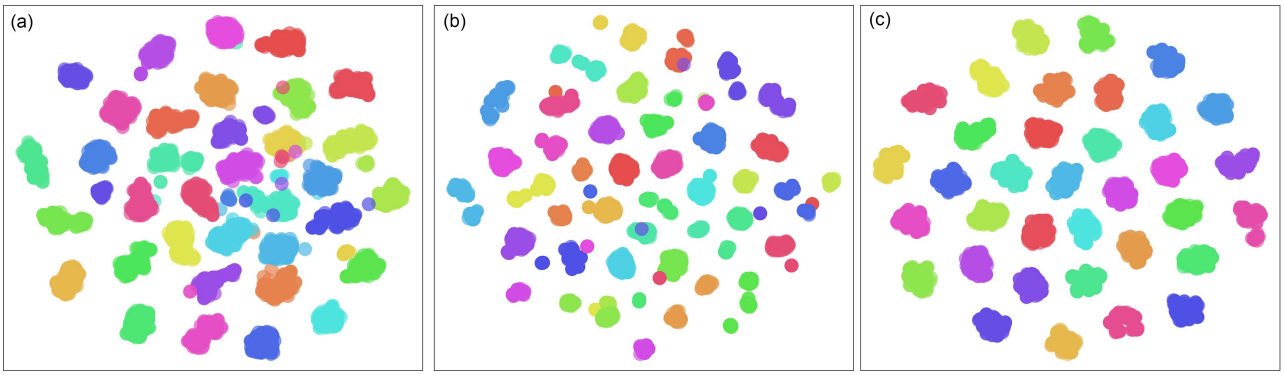

Figure 7 (Color online) t-SNE analysis of the feature representations. (a) Single-view representations with CSI amplitudes; (b) single-view representations with CSI phases; (c) dual-view representations.

Furthermore, at the 80th percentile accuracy, DVCLoc reduces the localization error by 20.4% for WiDeep, 24.1% for CNNLoc, 24.1% for GNN-Loc, 16% for DeepSense, 15.3% for DeepFi, and 14.9% for ReNet-Loc, respectively. Compared with traditional supervised methods, the proposed DVCLoc system has learned inherently view-invariant representations, thereby achieving greater environmental adaptability. DVCLoc also achieves higher accuracy than other solutions, even with the robustness of environmental dynamics.

5.3 Effect analysis and discussion

5.3.1 *Single-view vs. dual-view representations*

To evaluate the DVCLoc prototype's representation capability, we feed only amplitude or phase images into our pretraining model. As illustrated in Figure 7, t-SNE [35] analysis is employed to visualize the representation distributions for the single-view and dual-view images. Different colors indicate various distributions of the learned feature representations. DVCLoc can realize more distinct clusters with dual-view contrastive learning. We further quantify the view invariance to evaluate the quality of the learned representations in the DVCLoc system. The alignment metric γ_{dual} is defined as the expected distance between the feature vectors of dual-view sampling [36] as

$$\gamma_{\text{dual}} \triangleq \mathbb{E}\left[\|f(\mathbf{x}^a) - f(\mathbf{x}^p)\|_2^2\right]. \tag{21}$$

We calculate the alignment metric of both dual-view and single-view samplings with the amplitudes and the calibrated phase responses, such as $\gamma_{\text{dual}} = 0.52$, $\gamma_{\text{am}} = 0.58$, and $\gamma_{\text{ph}} = 0.61$. The experimental results indicate that the proposed DVCLoc system has better view-invariant quality than the other schemes, with the lowest value.

We also compare the RMSEs of the DVCLoc prototype with the single-view representations in Figure 8. By the inputs of the amplitude images, DVCLoc's localization accuracy improves by 18.68%, 6.82%, and 4.27% in the corridor, hall, and lounge, respectively. Meanwhile, feeding the calibrated phase images into our developed pretraining model, the RMSE of the proposed DVCLoc system reduces by 12.5%, 11.93%, and 8.2% in the corridor, hall, and lounge, respectively. From these localization results, DVCLoc effectively improves accuracy by mining the invariant factor underlying CSI data with intrinsic dual-view samplings.

5.3.2 *Effect on deployment density*

We evaluate the effect of the proposed DVCLoc system with RP deployment densities. For the downstream localization tasks, we randomly select different numbers of RPs to collect CSI fingerprints. The linear location predictor is trained in each specific indoor scenario. The deployment density is defined as $\rho = \frac{N_{lim}}{N_{all}}$, where N_{lim} is the number of the selected RP and N_{all} is the total number of RP in the testbed indoor scenario. With a range of the deployed density as $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, the RMSEs of location estimations are calculated in different indoor scenarios. As demonstrated in Figure 9, the localization errors gradually decrease with the increasing ρ, while the collection of numerous CSI fingerprints is labor-intensive. From the experimental results, the optimal tradeoff between deployment efforts and localization accuracy occurs at a deployed density of $\rho = 0.5$, indicating that DVCLoc achieves reasonable accuracy with half the predefined RP deployment.

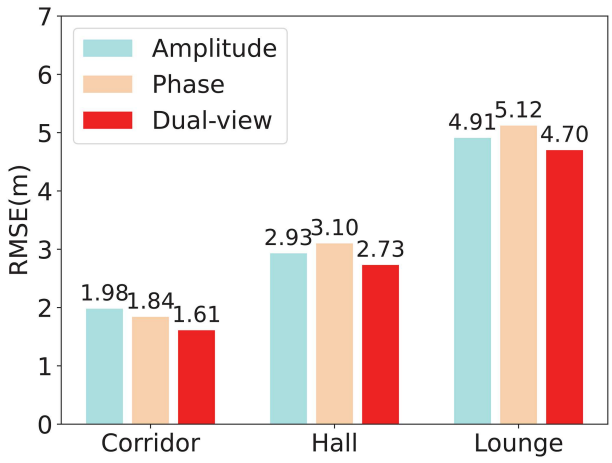

Figure 8 (Color online) Accuracy comparison with single-view and dual-view representations.

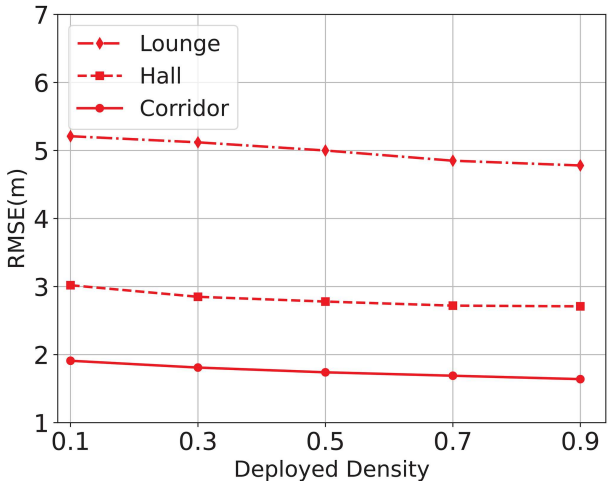

Figure 9 (Color online) Localization accuracy with different deployment densities.

5.4 Effect and effectiveness analysis of the dual-view contrastive pretraining model

We also evaluate the localization accuracy of the proposed DVCLoc system with the ablations of the dual-view contrastive pretraining model.

5.4.1 *Batch size*

We calculate the RMSE for the different batch sizes $\{128, 256, 512, 1024, 2048, 4096\}$. Figure 10(a) shows the localization results in diverse indoor scenarios. When the batch size is small, the learned representations across different views quickly converge due to gradient instability, leading to model collapse. When the batch size is 1024, DVCLoc achieves better localization accuracy than the other schemes do. However, increasing the batch size to 2048 or 4096 results in accuracy degradation, which may be caused by gradient randomness and redundant information in radio signal measurements.

5.4.2 *Learning rate*

The localization errors of DVCLoc are shown in Figure 10(b) for different values of the initial learning rate to update the learnable parameters of our pretraining model. DVCLoc requires an appropriate value to achieve satisfactory accuracy in diverse indoor scenarios. From the experimental results, DVCLoc's localization accuracy decreases as the learning rate increases. The best value for our dual-view contrastive pretraining model is 0.05.

5.4.3 *Low-dimensional projection*

We compute the RMSEs of downstream localization tasks in three typical indoor scenarios with and without (w/o) the low-dimensional projection. The experimental results are shown in Figure 10(c). With the low-dimensional projection, the localization accuracy is improved by 26.81%, 31.75%, and 32.85% in the corridor, hall, and lounge, respectively. The learnable projection networks are useful for our dual-view contrastive pretraining model to achieve high-quality feature representations and improve accuracy, especially in complex indoor scenarios.

5.4.4 *Feature dimension*

In the DVCLoc prototype, we evaluate the effect of the different dimensions on the feature embeddings from dual-view contrastive learning. As shown in Figure 10(d), the better localization accuracy is achieved with the feature dimension of 30 than the others. However, the increasing dimensionality of feature embeddings degrades localization accuracy, potentially leading to the learning of useless information.

5.4.5 *Dual-view contrastive loss*

With dual-view contrastive learning, we compare the localization accuracy with the symmetric and asymmetric contrastive losses in Figure 10(e). The localization errors of the symmetric contrastive loss are lower than those of

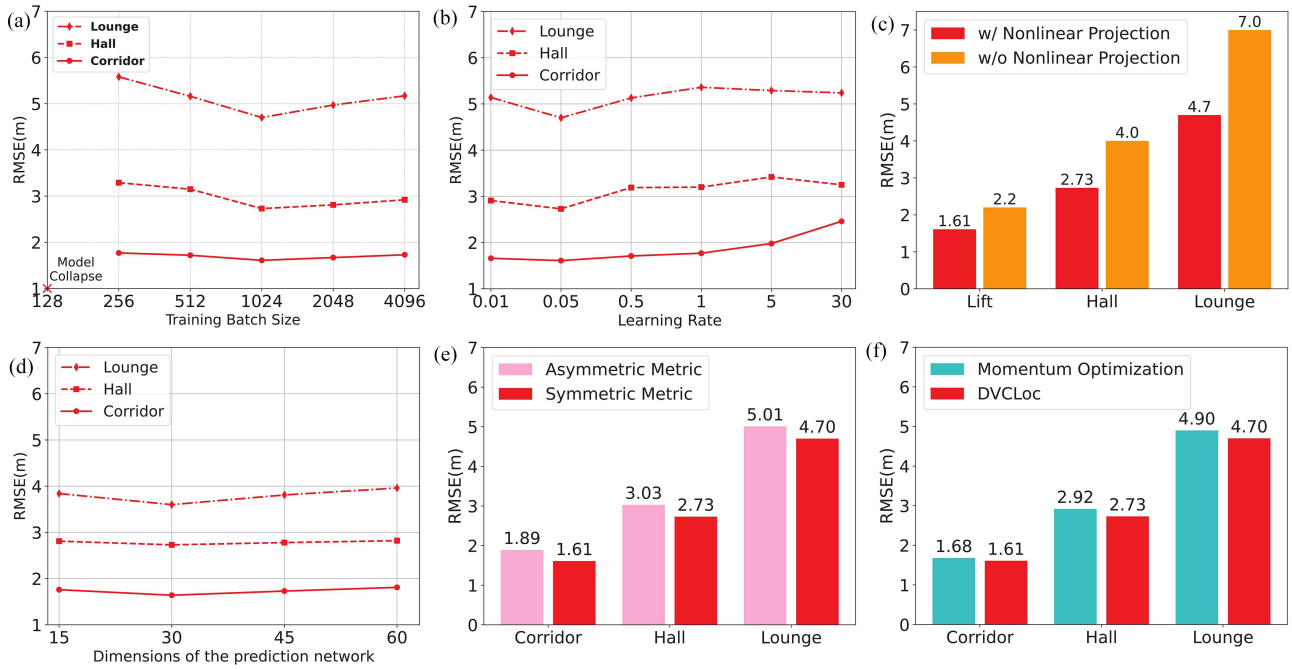

Figure 10 (Color online) Effects and effectiveness analysis of the proposed dual-view contrastive pretraining model. (a) Batch size; (b) learning rate; (c) low-dimensional projection; (d) feature dimension; (e) symmetric contrastive loss; (f) training strategy.

the asymmetric one by 14.8%, 9.9%, and 6.18% in the corridor, hall, and lounge, respectively.

5.4.6 *Training strategy*

In the framework of dual-view contrastive learning, we adopt the stop-gradient to avoid model collapse for learning view-invariant representations underlying dual-view samplings. We evaluate the RMSEs of localization estimations and compare this training strategy with the momentum optimization [30] in Figure 10(f). The DVCLoc prototype slightly improves localization accuracy.

6 Conclusion

In this paper, we propose a novel deep-learning-based localization framework to achieve ubiquitous location estimations with limited human effort. By refining the invariant factor underlying unannotated CSI data, the dual-view contrastive pretraining model is established to learn the view-invariant representations through the feature similarity maximization in a low-dimensional space. The learned feature encoders are transferable to capture the features in new environments or new scenarios. For robust localization, the location predictor is trained with limited CSI fingerprints, and accurate location estimation is achieved across various complex scenarios. Extensive experimental evaluation demonstrates that DVCLoc can achieve SoTA performance, effectively facilitating DNN-based localization for offering ubiquitous ILBSs. In the future, we will further fully utilize the multi-antenna and multi-subcarrier CSI data to expand the additional views of radio sampling for attaining performance improvement.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant No. 62171160), Natural Science Foundation of Heilongjiang Province (Grant No. JJ2025PL0282), Guangdong Provincial Key Laboratory (2024) (Grant No. 2024KSYS023), and Shenzhen Science and Technology Program (Grant No. JCYJ20241202124304007).

References

1 Kato N, Mao B, Tang F, et al. Ten challenges in advancing machine learning technologies toward 6G. IEEE Wireless Commun, 2020, 27: 96–103
2 Behravan A, Yajnanarayana V, Keskin M F, et al. Positioning and sensing in 6G: gaps, challenges, and opportunities. IEEE Veh Technol Mag, 2023, 18: 40–48
3 Fang K, Chen J X, Zhu H, et al. Explainable-AI-based two-stage solution for WSN object localization using zero-touch mobile transceivers. Sci China Inf Sci, 2024, 67: 170302
4 Zhu X, Qu W, Qiu T, et al. Indoor intelligent fingerprint-based localization: principles, approaches and challenges. IEEE Commun Surv Tut, 2020, 22: 2634–2657
5 He S, Chan S H G. Wi-Fi fingerprint-based indoor positioning: recent advances and comparisons. IEEE Commun Surv Tut, 2017, 18: 466–490

6 Vasisht D, Kumar S, Katabi D. Decimeter-level localization with a single WiFi access point. In: Proceedings of the 13th USENIX Conference on Networked Systems Design and Implementation, 2016. 165–178
7 Xiong J, Jamieson K. ArrayTrack: a fine-grained indoor location system. In: Proceedings of the 10th USENIX Conference on Networked Systems Design and Implementation, 2013. 71–84
8 Zhang L, Wang H. 3D-WiFi: 3D localization with commodity WiFi. IEEE Sens J, 2019, 19: 5141–5152
9 Ni J, Zhang F, Xiong J, et al. Experience: pushing indoor localization from the laboratory to the wild. In: Proceedings of the 28th Annual International Conference on Mobile Computing and Networking, 2022
10 Lymberopoulos D, Liu J. The microsoft indoor localization competition: Experiences and lessons learned. IEEE Signal Process Mag, 2017, 34: 125–140
11 Wang X, Gao L, Mao S, et al. CSI-based fingerprinting for indoor localization: a deep learning approach. IEEE Trans Veh Technol, 2017, 66: 763–776
12 Abbas M, Elhamshary M, Rizk H, et al. WiDeep: WiFi-based accurate and robust indoor localization system using deep learning. In: Proceedings of IEEE International Conference on Pervasive Computing and Communications (PerCom), 2019. 1–10
13 Song X, Fan X, Xiang C, et al. A novel convolutional neural network based indoor localization framework with WiFi fingerprinting. IEEE Access, 2019, 7: 110
14 Song R, Zhang D, Wu Z, et al. RF-URL: unsupervised representation learning for RF sensing. In: Proceedings of the 28th Annual International Conference on Mobile Computing And Networking, 2022. 282–295
15 Tong X, Wan Y, Li Q, et al. CSI fingerprinting localization with low human efforts. IEEE ACM Trans Netw, 2021, 29: 372–385
16 Wu C, Yang Z, Xiao C. Automatic radio map adaptation for indoor localization using smartphones. IEEE Trans Mobile Comput, 2018, 17: 517–528
17 Zhang L, Wu S, Zhang T, et al. Learning to locate: adaptive fingerprint-based localization with few-shot relation learning in dynamic indoor environments. IEEE Trans Wireless Commun, 2023, 22: 5253–5264
18 Li D, Xu J, Yang Z, et al. Train once, locate anytime for anyone: adversarial learning based wireless localization. In: Proceedings of IEEE Conference on Computer Communications, 2021. 1–10
19 Chen Y, Liu Z Y, Jiang F, et al. Localizing base stations with measured data: a concatenated image-based deep learning approach. Sci China Inf Sci, 2025, 68: 129301
20 Gao J, Wu D, Yin F, et al. MetaLoc: learning to learn wireless localization. IEEE J Sel Areas Commun, 2023, 41: 3831–3847
21 Xie Y, Li Z, Li M. Precise power delay profiling with commodity WiFi. In: Proceedings of the 21st Annual International Conference on Mobile Computing and Networking, 2015. 53–64
22 Gjengset J, Xiong J, McPhillips G, et al. Phaser: enabling phased array signal processing on commodity WiFi access points. In: Proceedings of the 20th Annual International Conference on Mobile Computing and Networking, 2014. 153–164
23 Shu Y, Huang Y, Zhang J, et al. Gradient-based fingerprinting for indoor localization and tracking. IEEE Trans Ind Electron, 2016, 63: 2424–2433
24 Yang X, Zhuang Y, Gu F, et al. DeepWiPos: a deep learning-based wireless positioning framework to address fingerprint instability. IEEE Trans Veh Technol, 2023, 72: 8018–8034
25 Zhang L, Qiu Y, Wu S, et al. GraphLoc: enhancing fingerprint-based localization with graph representation learning. IEEE Int Things J, 2025, 12: 21593–21603
26 Li Q, Liao X, Liu M, et al. Indoor localization based on CSI fingerprint by siamese convolution neural network. IEEE Trans Veh Technol, 2021, 70: 12168–12173
27 Zhang L, Wu S, Zhang T, et al. RobLoc: robust wireless localization with dynamic self-adaptive learning. IEEE Int Things J, 2024, 11: 17866–17877
28 Zhang L, Wu S, Zhang T, et al. Automatic radio map adaptation for robust indoor localization with dynamic adversarial learning. IEEE Trans Ind Inf, 2025, 21: 1615–1624
29 Kotaru M, Joshi K, Bharadia D, et al. SpotFi: decimeter level localization using WiFi. In: Proceedings of the ACM Conference on Special Interest Group on Data Communication, 2015. 269–282
30 Grill J B, Strub F, Altché F, et al. Bootstrap your own latent: a new approach to self-supervised learning. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 21271–21284
31 Chen X, He K. Exploring simple Siamese representation learning. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 15745–15753
32 Halperin D, Hu W, Sheth A, et al. Tool release: gathering 802.11n traces with channel state information. ACM SIGCOMM Comput Commun Rev, 2011, 41: 53–53
33 Wang S, Zhang S, Ma J, et al. Graph-neural-network-based WiFi indoor localization system with access point selection. IEEE Int Things J, 2024, 11: 33550–33564
34 Yao S, Hu S, Zhao Y, et al. Deepsense: a unified deep learning framework for time-series mobile sensing data processing. In: Proceedings of the 26th International Conference on World Wide Web, 2017. 351–360
35 Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res, 2008, 9: 2579–2605
36 Wang T, Isola P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: Proceedings of the 37th International Conference on Machine Learning, 2020. 9929–9939