remote sensing

MDPI

Article

Semi-Supervised FMCW Radar Hand Gesture Recognition via Pseudo-Label Consistency Learning

Yuhang Shi [1], Lihong Qiao [1,2,*], Yucheng Shu [1], Baobin Li [3], Bin Xiao [1], Weisheng Li [1] and Xinbo Gao [1]

[1] Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 2021211823@stu.cqupt.edu.cn (Y.S.); shuyc@cqupt.edu.cn (Y.S.); xiaobin@cqupt.edu.cn (B.X.); liws@cqupt.edu.cn (W.L.); gaoxb@cqupt.edu.cn (X.G.)
[2] Chongqing Big Data Collaborative Innovation Center, Chongqing 401135, China
[3] School of Information Science and Engineering, University of Chinese Academy of Sciences, Beijing 100871, China; libb@ucas.ac.cn
* Correspondence: qiaolh@cqupt.edu.com

Abstract: Hand gesture recognition is pivotal in facilitating human–machine interaction within the Internet of Things. Nevertheless, it encounters challenges, including labeling expenses and robustness. To tackle these issues, we propose a semi-supervised learning framework guided by pseudo-label consistency. This framework utilizes a dual-branch structure with a mean-teacher network. Within this setup, a global and locally guided self-supervised learning encoder acts as a feature extractor in a teacher–student network to efficiently extract features, maximizing data utilization to enhance feature representation. Additionally, we introduce a pseudo-label Consistency-Guided Mean-Teacher model, where simulated noise is incorporated to generate newly unlabeled samples for the teacher model before advancing to the subsequent stage. By enforcing consistency constraints between the outputs of the teacher and student models, we alleviate accuracy degradation resulting from individual differences and interference from other body parts, thereby bolstering the network's robustness. Ultimately, the teacher model undergoes refinement through exponential moving averages to achieve stable weights. We evaluate our semi-supervised method on two publicly available hand gesture datasets and compare it with several state-of-the-art fully-supervised algorithms. The results demonstrate the robustness of our method, achieving an accuracy rate exceeding 99% across both datasets.

Keywords: hand gesture recognition (HGR); semi-supervised learning; pseudo-label consistency; teacher–student model

check for updates

Citation: Shi, Y.; Qiao, L.; Shu, Y.; Li, B.; Xiao, B.; Li, W.; Gao, X. Semi-Supervised FMCW Radar Hand Gesture Recognition via Pseudo-Label Consistency Learning. *Remote Sens.* **2024**, *16*, 2267. https://doi.org/10.3390/rs16132267

Academic Editor: Stefano Tebaldini

Received: 13 May 2024
Revised: 8 June 2024
Accepted: 20 June 2024
Published: 21 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Hand gesture recognition (HGR) represents a research hotspot within the field of the Internet of Things (IoT) and is regarded as a means of human–machine interaction [1–4]. It allows us to communicate with machines in a more natural and convenient way. For disabled people, sign language serves as a means of communication. HGR technologies play a crucial role in facilitating smoother interactions between these individuals and the broader population.

Nowadays, there are various devices for capturing motion features. Based on their functionalities, data collection devices can be categorized into wearable devices, visual devices, and radar devices. Wearable devices require individuals to wear specific equipment on their bodies to collect signals, such as Surface Electromyogram (sEMG) [5]. While they can accurately capture gesture signals, wearable devices require continuous wearing throughout the measurement process, and the wearing process itself is complex. Visual devices typically rely on cameras and are commonly used in hand gesture recognition

due to their accessibility. However, they still face challenges associated with high computation costs and privacy leakage. Methods based on visual devices are sensitive to light conditions [6]. In contrast to visual devices, radar devices are unaffected by light conditions, enabling them to operate seamlessly at any time. High-resolution radar can effectively detect the tiny motions of fingers and muscles during hand movements [7]. Radars commonly used in gesture recognition include Frequency Modulated Continuous-Wave (FMCW) radar and Ultra-Wideband (UWB) radar. However, FMCW radar usually has higher accuracy than UWB radar. FMCW radar operates by coherently mixing a portion of the echo signal with the transmitted signal, resulting in an intermediate frequency signal that contains distance and velocity information of the target. Subsequent detection of the IF signal allows the extraction of motion information for the target. Because of its small antenna size and contactless operation capability, FMCW radar finds applications in various scenarios, including autonomous driving [8], sign language recognition [9–12], home automation [13,14], and many other fields. Therefore, FMCW radar plays a crucial role in human–machine interaction.

Regarding a typical recognition algorithm like hand gesture recognition, fully supervised methods are often employed. Before modeling, raw signals usually require preprocessing. For deep learning methods, tensor decomposition techniques [15,16] are often employed. These techniques decompose high-dimensional data into lower-dimensional representations (e.g., distance and velocity), which helps the model explicitly learn the relationships between signals and gestures. For example, Xia et al. [17] introduced a model called the moving scattering center model, which utilizes raw multi-channel data to derive a high-dimensional feature representation. This model extracts information about the range, angle, and velocity of the gesture from the multi-channel data. The recognition algorithm imposes constraints on the consistency of location, velocity and velocity changes. Furthermore, the algorithm utilizes a CNN model to categorize hand gestures. Dong et al. [18] introduced a method utilizing 3D fast Fourier transform, capable of processing dynamic hand gesture feature information in five dimensions simultaneously. The network comprises a jointly spatial and temporal deformable convolution block, along with an adaptable spatial and temporal hybrid convolution block that is sensitive to context. It can take full advantage of raw data and achieve an accuracy rate of 97.22%. However, acquiring such a scale of data is unaffordable. To reduce dependence on labeled data, Zhang et al. [19] proposed an unsupervised convolutional autoencoder network. By utilizing VGGNet, the input raw radar data representing target distances can be transformed into a latent space for feature representation. Then, a decoder is used to revert the latent representation into reconstructed signals. The analysis confirms that the autoencoder is capable of a 95% success rate in accuracy. While labels are not required during the training of the auto-encoder for feature extraction, fully supervised learning is still employed to train its classification capabilities. Additionally, Shen et al. [20] introduced a meta-learning network for rapid adaptation to unfamiliar hand gesture tasks with minimal training data. It is based on a 3D CNN fusion network which can leverage correlations in radar echo signal features. The extracted features are sent to a learnable relation module to classify and achieve an accuracy rate of 98.4%. It is worth noting that all of them predominantly rely on fully supervised methods. Despite attempts to reduce dependence on labeled data, as far as we know, there are currently no semi-supervised approaches for hand gesture recognition.

As a means of interaction, HGR algorithms have to overcome several challenges. Since variations exist in individual behaviors, there are some distinctions in dynamic hand gestures among different individuals. As a consequence, the same algorithm may exhibit varying performance levels under different individuals. It will be a challenge to the robustness of the algorithm. Methods of addressing robustness are typically divided into two categories: efficient algorithms and high-quality data. The former can be achieved through well-designed feature representation to enhance the ability of feature representation, while the latter can be attained by using high-resolution hardware. For traditional methods, they reach high accuracy through feature engineering, also resulting in robustness. Fan et al. [21]

introduced an approach that leverages an arcsine-based algorithm and motion imaging techniques to extract and reconstruct hand gesture features on a two-dimensional plane using demodulated Doppler phase shifts for gesture classification. Lee et al. [22] proposed a method based on a Hidden Markov Model with an ergodic model to categorize the gesture using the threshold of the image. Nai et al. [23] exploited the micro-Doppler signature to extract Chebyshev moments from the cadence velocity diagram obtained from the short-time Fourier transform of each recorded signal. Li et al. [24] introduced an innovative approach for dynamic hand gesture recognition utilizing radar sensors. Their method leverages the sparsity of radar hand gesture signals in the spectral-temporal space, represented by a Gabor dictionary, to extract micro-Doppler features. These features are then processed through the orthogonal matching pursuit algorithm and subsequently used by the sorting algorithm to identify various radar hand gesture signals. This technique demonstrates superior performance compared to traditional principal component analysis methods, achieving recognition accuracy rates exceeding 90%. Wang et al. [25] proposed a two-stage classification model that significantly enhances the accuracy of gesture type recognition by incorporating constraint functions into the Dynamic Time Warping (DTW) algorithm. This system is adept at recognizing and profiling gestures with minimal reliance on positioning or environmental conditions, thanks to its multisensor approach that gathers comprehensive data. The system's effectiveness is validated by experimental results, which indicate an average accuracy rate of 93.5% for hand gesture type recognition. Although traditional methods achieved good results in terms of robustness, they heavily relied on carefully designed feature engineering tailored to the specific structural differences between hand gestures.

However, signals are not easily understood intuitively by people, therefore it is difficult to discern the differences between them. In contrast, deep learning methods offer an alternative by automatically extracting features through neural networks, which can further take full advantage of the information in the signals. For instance, Liu et al. [26] proposed a dual-flow fusion deformable residual network (DFDRN) to enhance the robustness of HGR sensing by incorporating more detailed gesture information through time-frequency maps and spectrum map videos. In [27], Zhang et al. proposed a recurrent 3D convolutional neural network that can recognize dynamic hand gestures. To continuously recognize the gestures from unprocessed gestures in unsegmented input streams, they use a connectionist temporal classification algorithm based on LSTM to train the network and they achieve an accuracy rate of 96%. Furthermore, Wang et al. [28] proposed an LSTM-based classifier to distinguish different gesture signals. However, since LSTM relies on data, its capability is contingent upon the quality of the dataset. Consequently, enhancing robustness through LSTM proves challenging. To overcome this, Wang et al. [29] proposed HandNet with the aim of reducing the effect caused by different individuals during the feature representation process. To address this issue, they employed a strategy called Stepped Data Augmentation to mitigate data bias by integrating incoherent summation and uncovering relationships between successive frames. The model was evaluated on both single individuals in various scenarios and multiple individuals in the same scenario, achieving accuracy rates of 80.55% and 88.9%, respectively. In conclusion, while conventional methods struggle with the intuitive understanding of signals, deep learning approaches provide promising solutions by automatically extracting features through neural networks, thereby enhancing robustness in hand gesture recognition.

Despite achieving high accuracy, previous work encounters two primary challenges: the prohibitive cost associated with obtaining labeled data for supervised learning and ensuring method robustness. In addressing the challenges of robustness previously outlined, our approach employs an effective feature encoder. The encoder incorporates the Lite Transformer and effectively captures global and local features from gesture signals. To ensure the fidelity of the encoder's representation, we impose explicit constraints through end-to-end self-supervised learning. Furthermore, individual differences introduce redundant information into the signals, resulting in reduced model robustness and noise in the RDM

map. To address this, we introduce the Pseudo-label Consistency-Guided Mean-Teacher model, manually injecting noise into the teacher branch, and enforcing consistency constraints between the teacher and student models through semi-supervised learning. This ensures accurate prediction and minimizes the impact of noise, thereby enhancing model robustness. Concretely, we introduce a semi-supervised learning framework guided by pseudo-label consistency, characterized by a dual-branch structure. It consists of two main stages: the Global and Local Guided Self-Supervised Auto-Encoder stage, and the Pseudo-label Consistency-guided Mean-Teacher learning stage. In the first stage, the global and local guided self-supervised learning encoder is trained as a feature encoder, maximizing the use of available data to enhance the feature representation of radar signals. In the next stage, we employ a semi-supervised approach to train a Pseudo-label Consistency-Guided Mean-Teacher model, which effectively handles the impact of background noise by utilizing guided loss based on pseudo-label consistency. Compared to fully supervised methods, it can leverage unlabeled data to provide additional information to improve the capability of the model. In our framework, the student model branch utilizes original data, while the teacher model branch leverages unlabeled data. Through data augmentation techniques, we introduce specific noise to generate a significant amount of unlabeled data. By enforcing consistency constraints between the outputs of the teacher and student models, we mitigate accuracy degradation caused by individual differences and interference from other body parts, thereby enhancing the network's robustness. Furthermore, the teacher model aggregates the current state with historical states over a period via an exponential moving average to obtain robust and stable weights. This framework utilizes an effectively pre-trained feature encoder and semi-supervised learning for training the entire model, enhancing robustness while reducing dependence on extensive labeled datasets.

Our contributions can be summarized as followings:

- We introduce the Global and Local Guided self-supervised learning encoder, designed as a feature encoder. This encoder incorporates the Lite Transformer within a teacher–student network to effectively capture global and local features from gesture signals. Through self-supervised training, explicit constraints are applied to its inputs and outputs to ensure the fidelity of the encoder's representation, thereby enhancing model robustness.
- With an improved feature representation in the encoder, we further introduce the Pseudo-label Consistency-Guided Mean-Teacher model. By employing data augmentation techniques, we deliberately introduce noise to generate a substantial volume of unlabeled data. Through semi-supervised learning, we enforce consistency constraints between the outputs of the teacher and student models, ensuring the model's ability to predict accurately. This approach helps mitigate accuracy degradation stemming from individual differences and interference from other body parts, thereby enhancing the network's robustness.
- We evaluate our method on publicly available datasets and compare it with several state-of-the-art fully supervised algorithms. The results demonstrate our method achieving a best accuracy rate of over 99% on both the soil dataset and the air-writing dataset.

2. Signal Preprocessing

In our research, we employ the range-doppler map (RDM) [30] to depict dynamic hand gesture recognition. This method has found widespread application in radar-based HGR [29,31] due to its ability to resolve multiple scattering centers. The echo signal can be expressed as:

$$s(f,n) = \sum_{i=0}^{I} A^{(i)} \cdot e^{j(2\pi k\tau^{(i)} f + 2\pi f_d^{(i)} s + \phi^{(i)})} \quad (1)$$

Here, f represents the index of sampling points within a single chirp, referred to as fast time, while s denotes the index of the chirp's period, known as slow time. I stands for the number of scattering centers, $A^{(i)}$ signifies the amplitude of the i-th scattering center, k is

the wave number, $\tau^{(i)}$ denotes the corresponding time delay, $f_d^{(i)}$ represents the Doppler frequency caused by the moving hand, and $\phi^{(i)}$ is a phase term.

Through discrete Fourier transformation, the consecutive signals of $s(f,s)$ can be converted into RDM along both the fast and slow times. This process can be represented as:

$$\begin{aligned} r(p,q) &= \sum_{v=1}^{V}(\sum_{t=1}^{U} s(f,s)\cdot e^{-j2\pi pf/U})e^{-j2\pi qv/V} \\ &= \sum_{i=0}^{I}\alpha^{(i)}\delta(p-\omega_f^{(i)}, q-\omega_s^{(i)})\cdot e^{j\phi^{(i)}} \end{aligned} \tag{2}$$

Here, $\delta(\cdot)$ represents the Dirac function, U denotes the number of sampling points in a chirp, and V is the number of consecutive chirps. $\omega_f^{(i)} = 2\pi k\tau^{(i)}$ corresponds to the i-th scattering center's range, while $\omega_s^{(i)} = 2\pi f_d^{(i)} = 4\pi v^{(i)}\cos\theta^{(i)}/\lambda$ signifies the Doppler item. The magnitude of signal $|r(p,q)|$ is considered as the RDM of a specific frame. Notably, after obtaining the RDM of each frame in a gesture, we need to convert the video into an image because our model can only accept two-dimensional data. Specifically, we flatten the image in each frame and concatenate them along the time axis to create a final image representation.

3. Pseudo-Label Consistency Guided Mean-Teacher Model

In supervised learning, models typically require a large amount of labeled data to model the underlying mappings between features and categories. However, acquiring such a large volume of data is often time-consuming. In previous research, this challenge has been addressed by designing efficient feature encoders, yet it still demands substantial data. To tackle this issue, we introduce a Pseudo-label Consistency-Guided Mean Teacher model. It comprises two branches with identical architectures: the student branch and the teacher branch. Our goal is to achieve semi-supervised learning while reducing dependency on labels and ensuring robustness.

As illustrated in Figure 1, our model consists of two main stages: the Global and Local guided Self-supervised Auto-Encoder stage and the Pseudo-Label Consistency-Guided Mean-Teacher learning stage. In the first stage, self-supervised learning is used to train a feature encoder with the global and local feature encoders. In the next stage, we present a pseudo-label Consistency-Guided Mean-Teacher model, which mitigates the impact of background noise and Pseudo-Label Consistency-Guided loss. This framework, referred to as the Pseudo-label Consistency-Guided Mean-Teacher model, employs an effectively pretrained feature encoder and semi-supervised learning for training the entire model. In the subsequent sections, we will provide a detailed description of both the attention-based encoder and the whole model.

3.1. Global and Local Guided Self-Supervised Encoder of Hand Gesture Features

In this section, we provide a detailed explanation of the structure of the student model. As illustrated in Figure 1a, our baseline consists of three main components: Lite Transformer block-based Global Feature Encoder (GFE), CNN-based Local Feature Encoder (LFE), and a feature decoder. Through the utilization of self-attention, the GFE and LFE collaborate to create a hybrid feature encoder capable of capturing both long and short-range dependencies. In the following part of this section, we provide a detailed explanation of the network structure.

3.1.1. Global and Local Feature Encoder

The hybrid feature encoder aims to extract features while avoid noise in signal. To achieve the goal of efficient feature extraction, we choose Lite Transformer [32] as GFE. Compared to the vanilla Vision Transformer (ViT) [33], LT-block can maintain performance while reduce computational costs. Meanwhile, the CNN branch beside it focuses on

extracting local feature which can form a complementary relationship with the LT-block. Given an image from samples $I \in \{I_k\}_{k=1}^{n}$, the process of extracting global feature F_1 and local feature F_1 can be formulated as:

$$F_1 = \Phi_{\mathrm{GFE}}(I), F_2 = \Phi_{\mathrm{LFE}}(I) \tag{3}$$

where $\Phi_{\mathrm{GFE}}(\cdot)$ and $\Phi_{\mathrm{LFE}}(\cdot)$ are represent GFE and LFE, respectively.

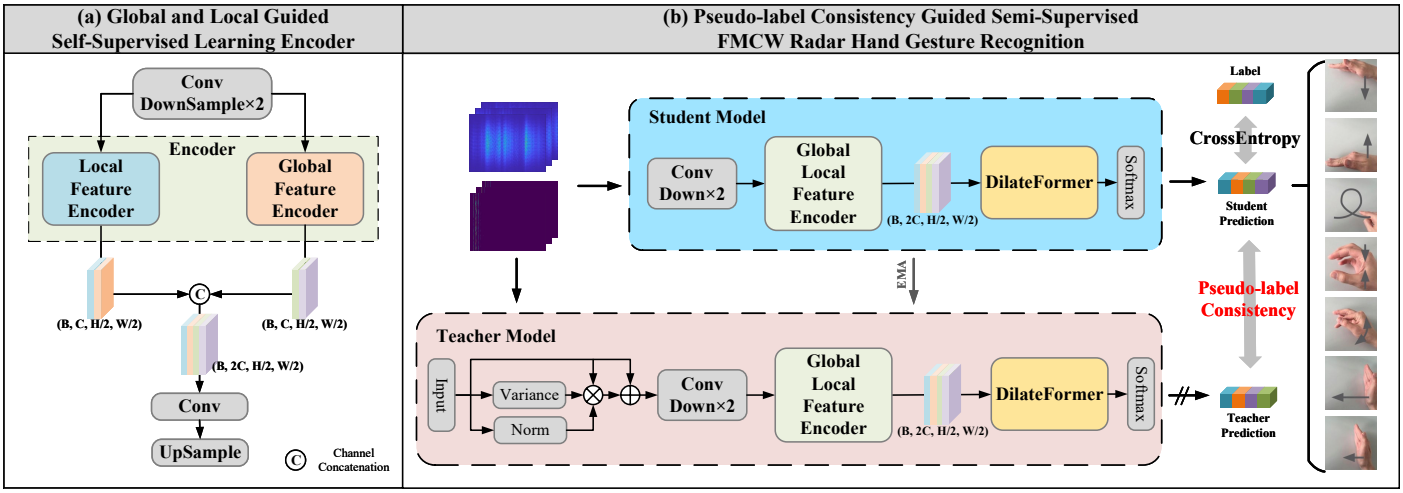


Figure 1. Overall pipeline of our suggested method. (**a**) Global and Local Guided Self-supervised Encoder (details in Section 3.1). The green component represents the encoder we proposed. The entire module in (**a**) processes gesture signals as input and generates reconstructed signals as output. (**b**) Pseudo-label Consistency Guided Semi-Supervised FMCW Radar Hand Gesture Recognition (details in Section 3.2). The blue and red parts are the mean-teacher network we proposed. The green component corresponds to the encoder from (**a**) and is initialized with the training parameters derived from (**a**), which keeps the robustness of the method. The pseudo-label consistency loss $\mathcal{L}_{con}$ is calculated based on the pseudo-labels generated by the teacher model from an augmented dataset and the predictions made by the student model. The update strategy for the teacher model involves utilizing the exponential moving average of the student model weights.

3.1.2. Global and Local Feature Decoder

To reconstruct the gesture signals, we employ a convolutional decoder to extract both global and local features from the input. Given the mismatch in size between the input and output signals, a two-scale upsampling module is employed to reconcile the disparity. During the mean-teacher learning phase, which resembles a typical recognition problem, we utilize a classifier in place of the global and local decoder. Considering the sparsity of signal features, computing long-range attention in a vanilla ViT consumes significant resources. Therefore, we employed the Dilateformer [34] to predict categories P for the hybrid features F, i.e.,

$$P = \Phi_{\mathrm{DF}}(F) \tag{4}$$

where Φ_{DF} represents DilateFormer. The reason we adopt DilateFormer is that it leverages Multi-Scale Dilated Attention to sparsely selected keys and values in sliding windows. By employing dilated attention, a balance can be struck between computational complexity and receptive field size.

3.1.3. Self-Supervised Training Loss

In the self-supervised training, the signal images I are initially fed into a feature encoder Φ_{in} for the extraction of shallow features. These shallow features serve as the foundation for the subsequent steps. To exploit these shallow features, the hybrid feature encoder $\{\Phi_{\mathrm{GFE}}, \Phi_{\mathrm{LFE}}\}$ use them to get global and local features, $\{F_1, F_2\}$, respectively. Once

we have obtained these output features, we merge them through channel-wise concatenation to create a unified set of hybrid features F. Afterward, the hybrid features F are directed to a decoder, denoted as Φ_{rec}, dedicated to feature reconstruction. Through the reconstruction decoder Φ_{rec}, we can reconstruct the hybrid features F to obtain the original signal images I_{rec}. The self-supervised loss can be formulated as:

$$\mathcal{L}_{self} = ||I - I_{rec}||_1. \tag{5}$$

Here, $\mathcal{L}_{self}$ represents the l_1-norm, while I and I_{rec} correspond to the original signal images and the reconstructed signal images, respectively.

3.2. Pseudo-Label Consistency-Guided Mean-Teacher Model for Hand Gesture Recognition

In this section, we provide a comprehensive overview of the mean-teacher model pipeline. As seen in Figure 1b, the model has two branches. Within the teacher branch, we introduce simulated noise to generate fresh samples for pseudo-labeling. Subsequently, we leverage the teacher model's predictions as pseudo-labels for training the student model. The learning process entails enforcing consistency constraints between predictions made on original samples and their augmented counterparts. This approach facilitates the model's adaptation to noisy data, thereby mitigating sensitivity to background noise and motion and ultimately enhancing overall robustness. Finally, weight updates are performed using the Exponential Moving Average. It is worth noting that the green encoder in Figure 1a is used to initialize the same encoder of the gesture signals in Figure 1b In the subsequent part of this section, we will offer a detailed explanation of our model.

3.2.1. Simulated Noise for Pseudo-Labeling Generation

In deep learning, extensive data is crucial for model training. However, constrained by radar equipment accessibility, it is challenging to gather a substantial amount of available data, resulting in a small-scale dataset that limits accuracy and robustness. In response to this, we introduce a simulated noise to generate new samples for pseudo-labeling. Concurrently, considering the difference between natural images and signal images, we select a kind of noise related to the basis of image attributes. So the augmentation does not result in a significant difference in similarity between the enhanced and original data. Concretely, for a normalized sample $I \in \{\mathrm{I}_k\}_{k=1}^{n}$, where $I \in R^{h \times w \times c}$. The standard deviation for each channel can be calculated as $\sigma \in \mathrm{R}^c$, where c is the index of channels. The initial noise is expressed as $n \sim N(0, \sigma^2 I)$, where σ^2, and I are element-wise multiplication and identity matrix, respectively. Through applying the softmax function, the pixel values of sample x of each channel can be re-normalized as $x_i{}'$. Ultimately, the noise η_{s} for augmentation can be formulated as:

$$n_{\mathrm{s}} = \lambda_s n(i, :, :, :) \cdot x_i{}' \cdot x_i \tag{6}$$

where λ_s is a hyper-parameter that controls the intensity of noise. The augmented image I_n can be written as $I_n = I + n_{\mathrm{s}}$.

3.2.2. Pseudo-Label Consistency-Guided Mean-Teacher Model for Semi-Supervised Gesture Recognition

During semi-supervised training, the encoder weights acquired through self-supervised training serve to initialize both the teacher and student models. We introduce a Pseudo-Label Consistency-Guided semi-supervised learning framework featuring a dual-branch structure. Specifically, our framework incorporates a mean-teacher network. Inspired by the mean teacher model [35,36], the student model is trained directly using the Adam optimizer to minimize the loss function, while the teacher model updates its parameters through EMA. Specifically, the weights of teacher model θ_t^{tea} at iteration t are derived from

the average of the teacher model weights θ_{t-1}^{tea} at iteration $t-1$ and the student model weights θ_t^{stu} at iteration t, i.e.,

$$\theta_t^{tea} = \alpha_t \theta_{t-1}^{tea} + (1-\alpha_t)\theta_t^{stu} \tag{7}$$

where α_t is a hyperparameter that balances the weight between the effects from history and current, representing the rate of weight decay, and its value at iteration t can be formulated as:

$$\alpha_t = \min(0.99, 1 - \frac{1}{1+t}) \tag{8}$$

This strategy helps the teacher model benefit from a longer memory. It allows the teacher model to produce more stable pseudo-labels while reducing the consistency gap between the teacher model and the student models, which can decrease the impact of inter-human differences on outcomes. Consequently, this mitigates the impact of inter-human differences on radar gesture recognition outcomes, ultimately improving the robustness of the model.

In semi-supervised training, the total semi-supervised loss becomes:

$$\mathcal{L}_{semi} = \mathcal{L}_{CE} + \mathcal{L}_{con} \tag{9}$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss utilized for training with labeled data, while $\mathcal{L}_{con}$ denotes the pseudo label consistency guided loss employed to enforce similarity between predictions generated by the teacher branch and the student branch. The Kullback–Leibler (KL) divergence is employed as the measure for the consistency-guided loss.

Given an image from the sample set $I \in \{I_k\}_{k=1}^{n}$, the cross-entropy loss $\mathcal{L}_{CE}$ between student's prediction and ground-truth is formulated as follows:

$$\mathcal{L}_{CE} = \mathbb{E}_{I \sim p_{stu}(I)}[-\log p_{gt}(I)] \tag{10}$$

where $p_{stu}(I)$ represents the outputs of the student branch and $p_{gt}(I)$ represents the ground truth for the image. Additionally, the consistency guided loss between the pseudo-label and prediction of the teacher, $\mathcal{L}_{con}$, is given by:

$$\mathcal{L}_{con} = D_{KL}[p_{stu}(I)||p_{tea}(I)] \ = \mathbb{E}_{I \sim p_{stu}(I)}[\log \frac{p_{stu}(I)}{p_{tea}(I)}] \tag{11}$$

where $p_{tea}(I)$ represents the outputs of the teacher branch.

4. Experiment

This section elaborates on the implementation and configuration specifics of the HGR task network. Experimental validations are conducted to substantiate both the efficacy of our model and the rationale behind the chosen network architecture.

4.1. Implement Details

Our experiments were conducted on a machine equipped with an NVIDIA GeForce RTX 3090 GPU and an Intel i9-10920X CPU. During preprocessing, training samples were resized to 600 × 400 patches. The training process consisted of 50 epochs, divided into two stages with 20 epochs for the first stage and 30 epochs for the second stage. A batch size of 4 was employed. We utilized the Adam optimizer, initializing the learning rate to 0.0001 and applying exponential decay, where the learning rate decreased by a factor of 0.9 after each epoch. In the local feature encoder, we utilize three convolutional layers, each with a stride of 3 and padding of 1, followed by a GeLU activation. For the global encoder, we implement a single transformer layer. The network settings follow the default configuration of the tiny version of DilateFormer.

4.2. Soli Data Set

4.2.1. Details

The Soli dataset [37] is a publicly available dataset developed by Google ATAP, specifically designed for gesture recognition in mobile and sensor applications. It has 4 receiving antennas and two transmitting antennas. Soli radar employs Frequency-Modulated Continuous Wave (FMCW) radar as a sensing technology to detect dynamic gestures in the air. The radar operates at a center frequency of 60 GHz, providing a range resolution of approximately 2 cm. Its remarkable temporal resolution allows it to discern minute movements at exceptionally high repetition rates, ranging between 1–10 kHz. The radar system is equipped with four receiving antennas, resulting in four channels to represent a given gesture. Considering the inherent self-obstruction phenomenon of radar, it is observed that during signal transmission and reception, a portion of the radar echo becomes obstructed by either the body or background objects, thereby significantly influencing the radar's detection efficacy and the quality of acquired data. In order to optimize detection performance, hand movements are ideally directed towards the radar's orientation. Nevertheless, due to spatial constraints within the compact radar device, such parameters were regrettably overlooked in the dataset.

The Soli dataset comprises 11 unique dynamic gestures, as depicted in Figure 2, including gestures like Pinch index, Palm tilt, and Finger slider, among others. Detailed design instructions and considerations are outlined in [37]. This dataset is divided into two sections: cross-person and cross-scenario. The cross-person section consists of data collected from ten individuals under consistent conditions, totaling 2750 samples. Conversely, the cross-scenario subpart encompasses data gathered from a single individual across six distinct scenarios. In total, the dataset contains 5500 samples. Examples of various signal data samples utilized in our methodology are illustrated in Figure 3.

4.2.2. Results on Cross-Person Gesture Recognition

Firstly, we compare our method on Soli cross-person data with CNN-LSTM [37], a real-time method based on a stacked CNN module and LSTM module, Gesturevlad [38] based on CNN-LSTM and an unsupervised feature representation module, which can aggregate spatial and temporal features at once, and Handnet [29], which employs Stepped Data Augmentation to integrate incoherent summation and uncover relationships between successive frames. For comparison, we set the ratio of the training set to the test set to 0.8:0.2.

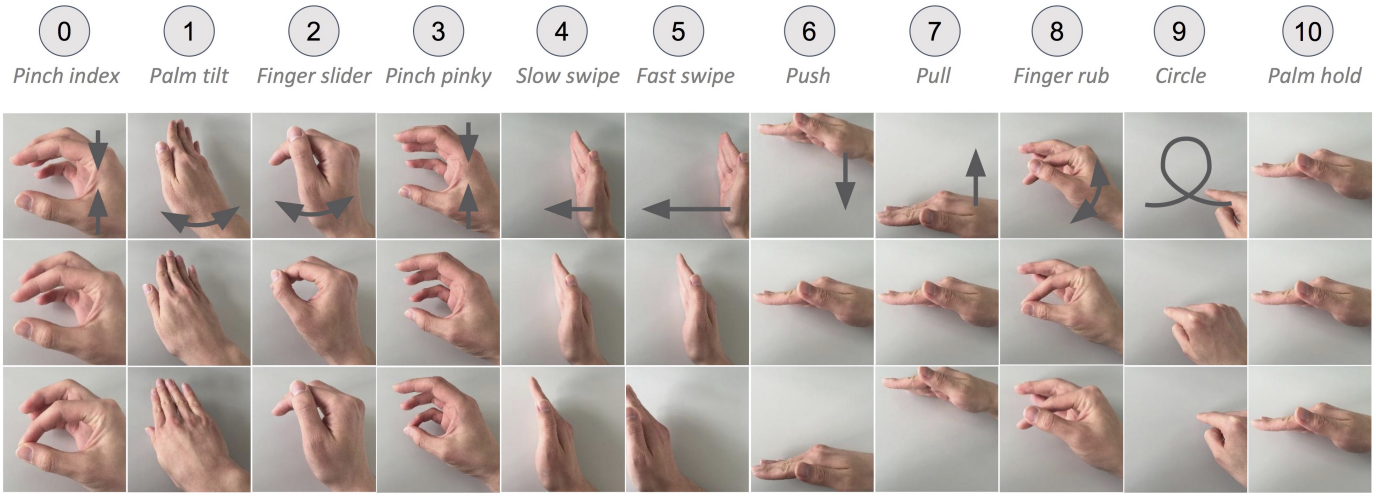


Figure 2. The eleven hand gestures in Soli dataset detected by FMCW radar. Arrows indicate the trajectory of the gesture motion.

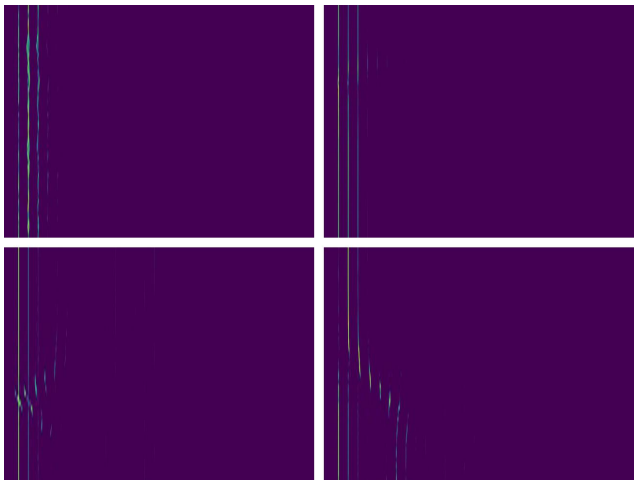


Figure 3. Four examples of model input signals in the Soli dataset.

As shown in Table 1, we achieve an average accuracy of over 99.27%. Compared with previous work, our methodology leverages an efficient model grounded on both global and local guided feature encoders. The encoder can decouple local and global features to fully leverage the distinct representations among individual users. Through self-supervised training, we can ensure the learning features are accurately represented. Furthermore, employing a pseudo-label guided semi-supervised mean-teacher model helps mitigate disruptions from noise, resulting in stable model weights and yielding the best score on the test set.

Table 1. Comparison results of different methods in cross-person and cross-scenario gesture recognition scenarios. The best results are shown in **bold**.

Cross-Person	Avg. Acc.	G0	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Gesturevlad [38]	78	65	93.37	62.5	62.78	76.95	60.18	95.98	90.46	90.51	85.9	74.41
CNN-LSTM [37]	79.06	58.71	95.16	64.8	67.92	72.31	72.91	93.4	89.99	91.82	82.8	80.24
HandNet [29]	80.55	68.59	92.04	60.72	70.27	75.98	84	92.34	89.57	86.66	82.36	83.53
Our method	**99.27**	**100**	**100**	**100**	**100**	**92**	**100**	**100**	**100**	**100**	**100**	**100**
Cross-scenario	Avg. Acc.	G0	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Gesturevlad [38]	84.01	59.58	96.63	73.51	63.52	86.74	84.55	99.06	96.12	93.66	91	79.84
CNN-LSTM [37]	85.75	56.69	95.33	76.43	61.98	92.73	81.39	98.42	97.79	96.83	96.92	89.1
HandNet [29]	88.9	83.16	95.24	76.62	65.74	93.06	98.17	97.74	97.7	92.04	96.31	82.15
Our method	**99.81**	**100**	**100**	**98**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

The confusion matrix is illustrated in Figure 4. On the horizontal axis, we observe the ground-truth labels of the samples, while the vertical axis displays the predicted categories of these samples. Each intersection point reflects the model's prediction results for a particular true class of samples. Within the confusion matrix, the diagonal elements indicate the count of correct predictions, while the off-diagonal elements signify the count of incorrect predictions. Analyzing the confusion matrix, we note that our method achieves 100% average accuracy across all gestures, except for G4. Nonetheless, when juxtaposed with alternative methods, our inaccuracies are deemed acceptable.

To demonstrate the effectiveness of the recognition capability of our model, we calculate the F_1 score, a widely used metric for evaluating classification capabilities. Before computing it, we must first obtain the precision and recall scores. To calculate the precision score in multiclass classification, we designate one class as the positive class and consider the rest as negative classes. Precision is calculated using True Positives (TP) and False Positives (FP), where TP is the number of correctly predicted instances of the designated

class, and FP is the number of instances predicted as the designated class but actually belonging to other classes, as shown in the following formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

Notably, after obtaining scores for each class, averaging is required to derive the final precision score. Similarly, recall is calculated using True Positives and False Negatives (FN), where FN represents the number of instances of the designated class predicted as other classes. It can be formulated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

As depicted in the confusion matrix, the precision and recall scores are 0.9929 and 0.9927, respectively. By applying the following formula, we derive the F_1 score:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

The model's F_1 score is 0.9928, which showing the high capability of gesture recognition.

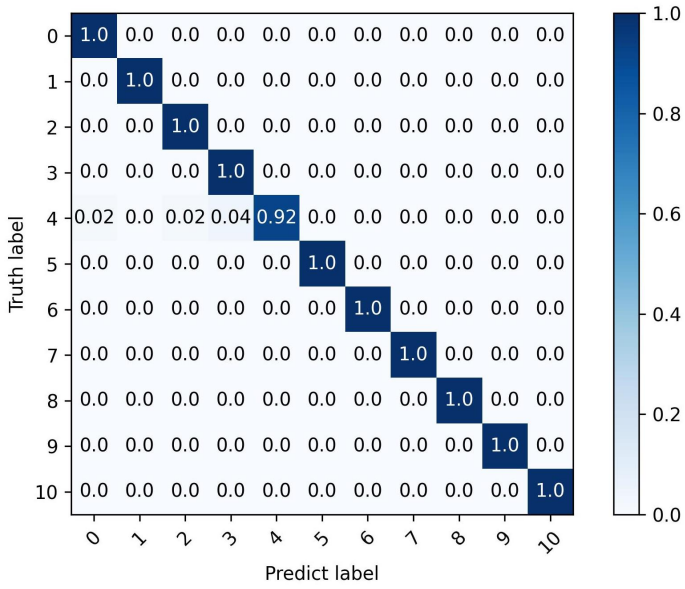


Figure 4. The confusion matrix results on the Soli cross-person test set.

The T-distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimensionality reduction algorithm extensively employed in data visualization. Renowned for its ability to preserve the underlying topological structure of the original data, t-SNE primarily serves as a tool for visualizing high-dimensional datasets. The algorithm operates by computing the similarities among data points in high-dimensional space and subsequently projecting these similarities onto a lower-dimensional space. In our study, we utilize the output of the final block from DilateFormer as the input high-dimensional features for visualizing the feature representation. The evaluation results obtained from the test data, illustrated in Figure 5, indicate that our method excels at distinguishing a significant portion of the features.

4.2.3. Results on Cross-Scenario Gesture Recognition

Following this, our methodology is implemented on a specific subset of data referred to as the Cross-scenario subset within the Soli dataset. This subset comprises gestures performed by a single individual across various scenarios. As demonstrated in Table 1, our

method underwent thorough evaluation within the Soli cross-scenario context, achieving an impressive average accuracy of 99.45%. The results indicate that our method exhibits robustness to environmental changes. In comparison to previously mentioned methods, our approach has attained state-of-the-art performance.

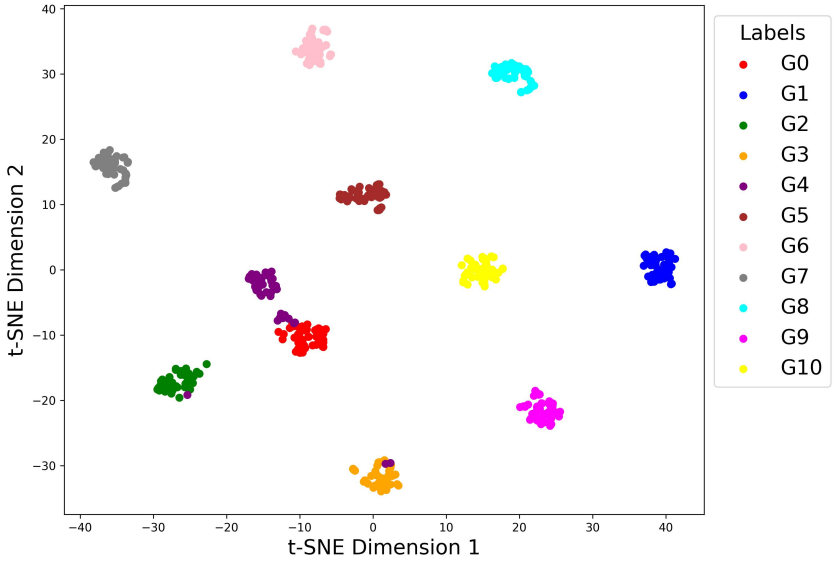


Figure 5. The feature distribution results obtained via t-SNE algorithm on Soli cross person test set.

The evaluation of the confusion matrix conducted on the cross-scenario subset is depicted in Figure 6. In this specific scenario, our method has exhibited outstanding performance, attaining 100% accuracy across all categories except for G2. Subsequently, we compute the Precision and Recall scores. Utilizing the aforementioned formulas, we obtain scores of 0.9982 and 0.9981, respectively. By incorporating precision and recall, we derive the F_1 score for our method on the Cross-scenario subset, yielding 0.9982, thereby showcasing the effectiveness of our approach. Additionally, employing t-SNE visualization techniques, we have provided a comprehensive depiction of the feature distribution of the classifier on the test set. The results, as presented in Figure 7, unequivocally reveal the existence of well-defined and discernible boundaries delineating each category of gesture features.

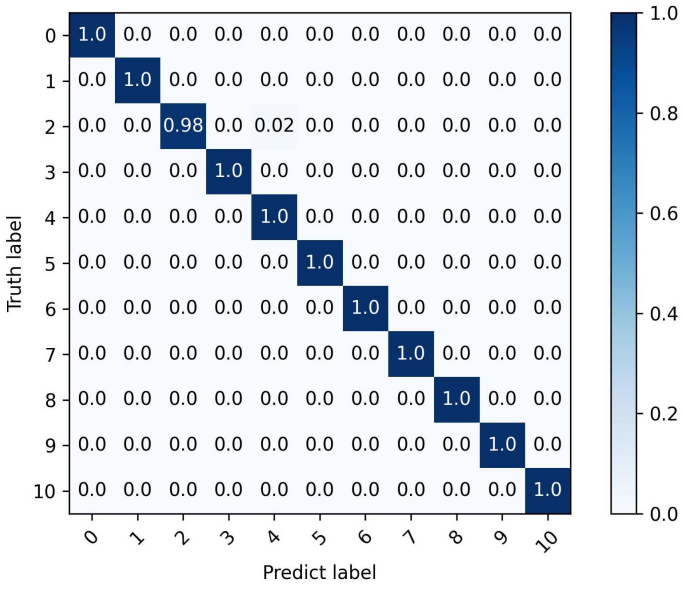


Figure 6. The confusion matrix results on the Soli cross scenario test set.

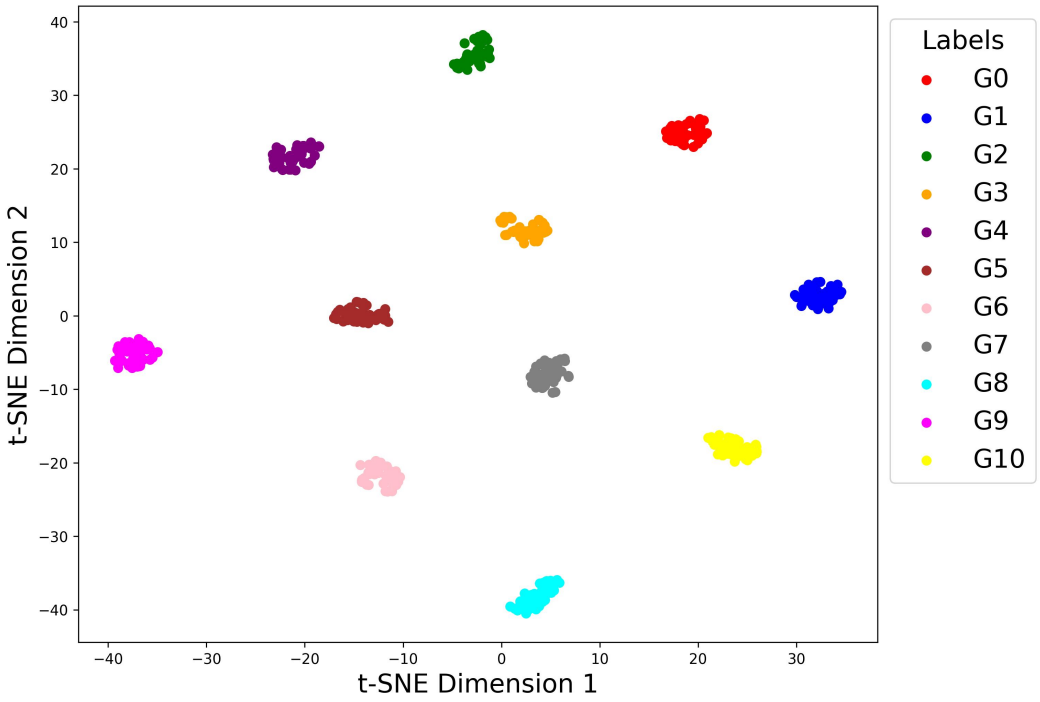


Figure 7. The feature distribution results obtained via t-SNE algorithm on Soli cross scenario test set.

4.3. Air-Writing Data Set

4.3.1. Detailed

The air-writing dataset [39] was generated utilizing the TI-IWR6843 Frequency Modulated Continuous Wave (FMCW) radar integrated with the TI-DCA1000 Field-Programmable Gate Array (FPGA) kit to capture dynamic gestures. The radar operates with four receive antennas and three transmit antennas, functioning at a central frequency of 62 GHz and a bandwidth of 4 GHz. As depicted in the fig, the radar was placed on the right side during the execution of gestures.

The air-writing dataset collecting settings, depicted in Figure 8a,b, features the radar positioned on the right side of the hand. It comprises two scenarios: home and laboratory, encompassing a total of ten gestures ranging from 0 to 9. The motion trajectories of each gesture are illustrated in Figure 8c. To enhance the robustness of the system in diverse environmental conditions, data were recorded at two distinct locations. The data collection process involved the participation of 12 individuals, with 100 air-writing samples acquired from each participant. Each air-writing sample consisted of 100 frames, collectively providing a comprehensive representation of the gesture dynamics. Several data samples are illustrated in Figure 9.



(**a**) (**b**) (**c**)

Figure 8. Radar setups and gesture capture areas in (**a**) home and (**b**) laboratory environments. Capturing (**c**) gesture trajectories for the digital numbers.

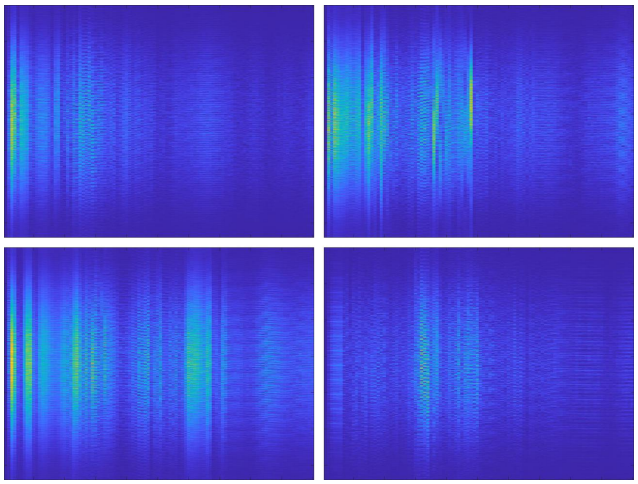


Figure 9. Four examples of model input signals in the Air writing dataset.

4.3.2. Results on Air Writing Gesture Recognition

When preprocessing the data prior to the training phase, several distinct steps were employed to convert the data matrix into an RGB image. Specifically, we normalized the matrix by subtracting its mean and applied a filter [40] to generate the image. Due to the unavailability of raw data, our method had to be evaluated on images obtained from the preceding steps. The experimental results demonstrate an average accuracy exceeding 99%. It is noteworthy that the data was partitioned into training and testing sets in the ratio of 0.8 and 0.2, respectively.

The confusion matrix for the test data analyzed in air writing is presented in Figure 10. The results show that our algorithm predicts well in each class. Consistent with the previously described procedure, the vertical axis represents the true labels, while the horizontal axis indicates the model's predicted labels. Impressively, aside from G8, we achieved a perfect accuracy rate across handwritten digits from 0 to 9. To further underscore the effectiveness of our method, precision, and recall scores were calculated from the confusion matrix, resulting in 0.9943 and 0.994, respectively. Consequently, the F_1 score is 0.9941. The experimental results show that the proposed model can distinguish the high-dimensional features of radar data under semi-supervised situation.

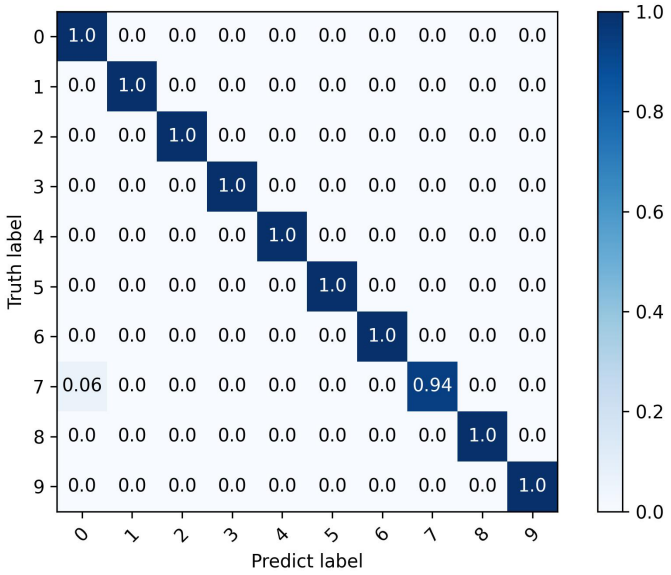


Figure 10. The confusion matrix results on the air writing test set.

Furthermore, employing the t-SNE technique, we performed dimensionality reduction on the test set, projecting high-dimensional features onto a lower-dimensional space to visualize their distribution within the coordinate system. The outcomes depicted in Figure 11 confirm the presence of distinct and discernible boundaries among each category of gesture features.

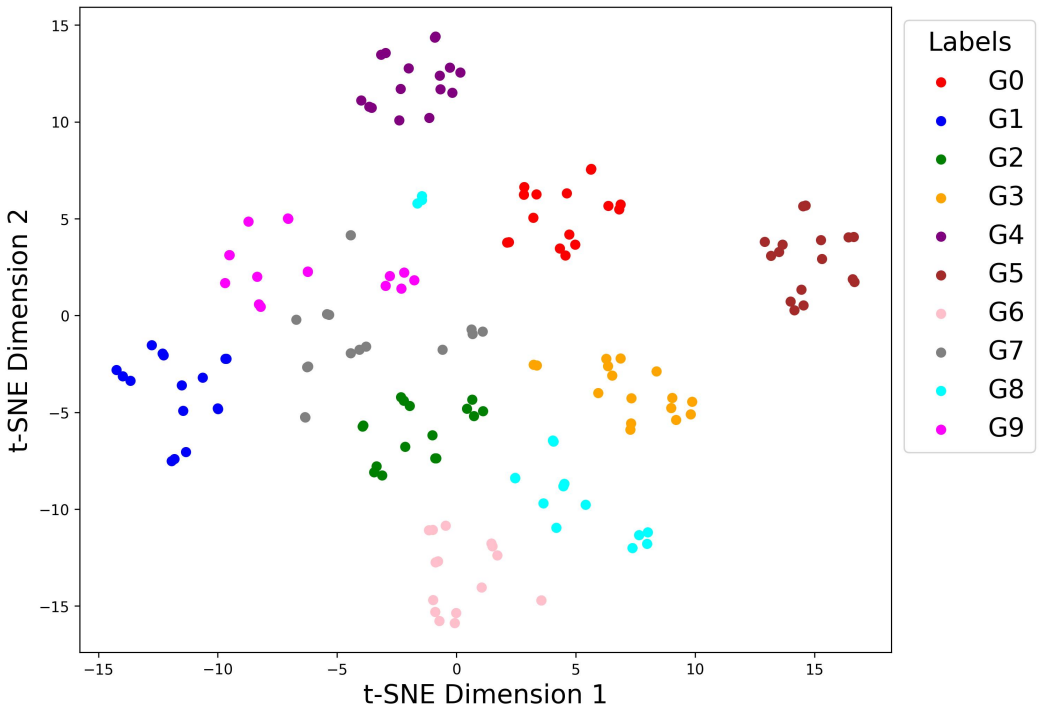


Figure 11. The feature distribution results obtained via t-SNE algorithm on air writing test set.

4.3.3. Ablation Study

We conducted tests on both the Soli dataset and air-writing datasets to validate the effectiveness of our various modules. To demonstrate the effectiveness of our feature representation, we replaced the global and local guided encoder with only the local feature encoder or global feature encoder. The results in Table 2 indicate that using a single encoder performs worse than the original configuration. Because the global and local encoder aims to extract the different types of features. In addition, we performed experiments on teacher models and self-supervised learning. The inclusion of an additional teacher model allows the student model to achieve more stable weights, thereby improving its performance on the test set. The self-supervised learning explicitly constrains the representation of signal features in the encoder. The accuracy values presented in the table illustrate the effectiveness of our mean-teacher model and the two-stage training strategy. In summary, ablation results underscore the efficacy and rationale behind our network design.

Table 2. Ablation experiment results in the test set of Soli and Air-writing. The best results are shown in **Bold**.

Configurations	Air-Writing	Cross-Scenario	Cross-Person
w/o local feature encoder	94.37	99.63	97.27
w/o global feature encoder	99.37	99.63	98.9
w/o teacher model	98.12	99.27	98.72
w/o self-supervised learning	98.75	99.63	97.63
Our method	**99.4**	**99.81**	**99.27**

5. Discussion

Hand gesture recognition has already found applications in various domains: it enables control of home appliances (e.g., lights) in smart homes, and interprets driving instructions through gestures in autonomous vehicles. In the next step, we aim to advance gesture recognition technology by focusing on more complex and dynamic real-world environments. Such applications require rapid and accurate gesture recognition in noisy and diverse settings, including scenarios involving multiple individuals or changing backgrounds. Our future research will concentrate on refining and enhancing existing gesture recognition algorithms to improve their accuracy and robustness. Our goal is to ensure that these systems can seamlessly adapt to various environmental and situational contexts, enabling more reliable and efficient real-world applications.

6. Conclusions

In this study, we propose a framework for gesture recognition to address issues related to labeling and robustness. Specifically, we present a semi-supervised learning framework guided by pseudo-label consistency, employing a dual-branch structure with a mean-teacher network. In our teacher–student network, the global and locally guided self-supervised learning encoder is utilized as the feature extractor to encode features efficiently, aiming to fully utilize existing data. Prior to the second stage, we introduce specific simulated noise to generate additional unlabeled samples for the teacher model. Through the constraint of consistency between the predictions and pseudo-labels from the student and teacher models, respectively, we enable the mitigation of accuracy decline caused by individual differences and interference from other body parts, thereby improving the robustness of the model. Finally, the teacher model is fine-tuned using Exponential Moving Average. In our experimental evaluations, we assess our approach on two publicly available datasets and compare it against several state-of-the-art algorithms. The results showcase the robustness of our method, achieving an accuracy rate exceeding 99% on both datasets.

Author Contributions: Y.S. (Yuhang Shi) conducted the experiment and wrote the paper. L.Q. and B.X. conceived original ideas and structured the paper. Y.S. (Yucheng Shu) and B.L. processed the raw data and analyzed the results. W.L. and X.G. revised the paper and supervised the project. All authors have read and agreed to the published version of the manuscript.

Funding: This work received partial support from various funding sources, including the National Key Research and Development Project, China (Grant 2019YFE0110800), the National Natural Science Foundation of China (Grants 62276040, 62276041, 62221005 and 61976031), the National Key Research Instrument Development Program, China (Grant 62027827), Chongqing Education Commission Science and Technology Research Project, China (Grant KJQN202200624), and Chongqing Big Data Collaborative Innovation Center Funding (Grant CQBDCIC202303).

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sahoo, J.P.; Sahoo, S.P.; Ari, S.; Patra, S.K. Hand Gesture Recognition Using Densely Connected Deep Residual Network and Channel Attention Module for Mobile Robot Control. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 5008011. [CrossRef]
2. Gurbuz, S.Z.; Amin, M.G. Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring. *IEEE Signal Process. Mag.* **2019**, *36*, 16–28. [CrossRef]
3. Ahmed, S.; Kallu, K.D.; Ahmed, S.; Cho, S.H. Hand gestures recognition using radar sensors for human-computer-interaction: A review. *Remote Sens.* **2021**, *13*, 527. [CrossRef]
4. Zhao, Y.; Liu, T.; Feng, X.; Zhao, Z.; Cui, W.; Fan, Y. New application: A hand air writing system based on radar dual view sequential feature fusion idea. *Remote Sens.* **2022**, *14*, 5177. [CrossRef]
5. Meng, L.; Jiang, X.; Liu, X.; Fan, J.; Ren, H.; Guo, Y.; Diao, H.; Wang, Z.; Chen, C.; Dai, C.; et al. User-tailored hand gesture recognition system for wearable prosthesis and armband based on surface electromyogram. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 2520616. [CrossRef]

6. Mitra, S.; Acharya, T. Gesture recognition: A survey. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2007**, *37*, 311–324. [CrossRef]
7. Lien, J.; Gillian, N.; Karagozler, M.E.; Amihood, P.; Schwesig, C.; Olson, E.; Raja, H.; Poupyrev, I. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Trans. Graph. (TOG)* **2016**, *35*, 1–19. [CrossRef]
8. Molchanov, P.; Gupta, S.; Kim, K.; Kautz, J. Hand gesture recognition with 3D convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 1–7.
9. Li, B.; Yang, Y.; Yang, L.; Fan, C. Sign Language/Gesture Recognition on OOD Target Domains Using UWB Radar. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2529711. [CrossRef]
10. Hein, Z.; Htoo, T.P.; Aye, B.; Htet, S.M.; Ye, K.Z. Leap motion based myanmar sign language recognition using machine learning. In Proceedings of the 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), St. Petersburg, Moscow, Russia, 26–29 January 2021; pp. 2304–2310.
11. Yang, Z.; Zhen, Z.; Li, Z.; Liu, X.; Yuan, B.; Zhang, Y. RF-CGR: Enable Chínese Character Gesture Recognition with RFID. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 8006116. [CrossRef]
12. Li, B.; Yang, Y.; Yang, L.; Fan, C. Objective Evaluation of Clutter Suppression for Micro-Doppler Spectrograms of Hand Gesture/Sign Language Based on Pseudo Reference Image. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5105113. [CrossRef]
13. Majdoub Bhiri, N.; Ameur, S.; Alouani, I.; Mahjoub, M.A.; Ben Khalifa, A. Hand gesture recognition with focus on leap motion: An overview, real world challenges and future directions. *Expert Syst. Appl.* **2023**, *226*, 120125. [CrossRef]
14. Rastgoo, R.; Kiani, K.; Escalera, S. Sign language recognition: A deep survey. *Expert Syst. Appl.* **2021**, *164*, 113794. [CrossRef]
15. Zhang, R.; Cheng, L.; Wang, S.; Lou, Y.; Gao, Y.; Wu, W.; Ng, D.W.K. Integrated Sensing and Communication with Massive MIMO: A Unified Tensor Approach for Channel and Target Parameter Estimation. *IEEE Trans. Wirel. Commun.* 2024, *early access*. [CrossRef]
16. Xie, R.; Hu, D.; Luo, K.; Jiang, T. Performance analysis of joint range-velocity estimator with 2D-MUSIC in OFDM radar. *IEEE Trans. Signal Process.* **2021**, *69*, 4787–4800. [CrossRef]
17. Xia, Z.; Luomei, Y.; Zhou, C.; Xu, F. Multidimensional feature representation and learning for robust hand-gesture recognition on commercial millimeter-wave radar. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4749–4764. [CrossRef]
18. Dong, X.; Zhao, Z.; Wang, Y.; Zeng, T.; Wang, J.; Sui, Y. FMCW radar-based hand gesture recognition using spatiotemporal deformable and context-aware convolutional 5-D feature representation. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5107011. [CrossRef]
19. Zhang, Z.; Tian, Z.; Zhang, Y.; Zhou, M.; Wang, B. u-DeepHand: FMCW radar-based unsupervised hand gesture feature learning using deep convolutional auto-encoder network. *IEEE Sens. J.* **2019**, *19*, 6811–6821. [CrossRef]
20. Shen, X.; Zheng, H.; Feng, X.; Hu, J. ML-HGR-Net: A meta-learning network for FMCW radar based hand gesture recognition. *IEEE Sens. J.* **2022**, *22*, 10808–10817. [CrossRef]
21. Fan, T.; Ma, C.; Gu, Z.; Lv, Q.; Chen, J.; Ye, D.; Huangfu, J.; Sun, Y.; Li, C.; Ran, L. Wireless hand gesture recognition based on continuous-wave Doppler radar sensors. *IEEE Trans. Microw. Theory Tech.* **2016**, *64*, 4012–4020. [CrossRef]
22. Lee, H.K.; Kim, J.H. An HMM-based threshold model approach for gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 961–973.
23. Nai, W.; Liu, Y.; Rempel, D.; Wang, Y. Fast hand posture classification using depth features extracted from random line segments. *Pattern Recognit.* **2017**, *65*, 1–10. [CrossRef]
24. Li, G.; Zhang, R.; Ritchie, M.; Griffiths, H. Sparsity-based dynamic hand gesture recognition using micro-Doppler signatures. In Proceedings of the 2017 IEEE Radar Conference (RadarConf), Seattle, WA, USA, 8–12 May 2017; pp. 0928–0931.
25. Wang, Z.; Yu, Z.; Lou, X.; Guo, B.; Chen, L. Gesture-radar: A dual doppler radar based system for robust recognition and quantitative profiling of human gestures. *IEEE Trans. Hum.-Mach. Syst.* **2020**, *51*, 32–43. [CrossRef]
26. Liu, Z.; Liu, H.; Ma, C. A robust hand gesture sensing and recognition based on dual-flow fusion with FMCW radar. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4028105. [CrossRef]
27. Zhang, Z.; Tian, Z.; Zhou, M. Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor. *IEEE Sens. J.* **2018**, *18*, 3278–3289. [CrossRef]
28. Wang, Y.; Wang, D.; Fu, Y.; Yao, D.; Xie, L.; Zhou, M. Multi-hand gesture recognition using automotive FMCW radar sensor. *Remote Sens.* **2022**, *14*, 2374. [CrossRef]
29. Wang, L.; Cui, Z.; Pi, Y.; Cao, C.; Cao, Z. Low personality-sensitive feature learning for radar-based gesture recognition. *Neurocomputing* **2022**, *493*, 373–384. [CrossRef]
30. Molchanov, P.; Gupta, S.; Kim, K.; Pulli, K. Short-range FMCW monopulse radar for hand-gesture sensing. In Proceedings of the 2015 IEEE Radar Conference (RadarCon), Arlington, VA, USA, 10–15 May 2015; pp. 1491–1496.
31. Min, R.; Wang, X.; Zou, J.; Gao, J.; Wang, L.; Cao, Z. Early gesture recognition with reliable accuracy based on high-resolution IoT radar sensors. *IEEE Internet Things J.* **2021**, *8*, 15396–15406. [CrossRef]
32. Wu, Z.; Liu, Z.; Lin, J.; Lin, Y.; Han, S. Lite transformer with long-short range attention. *arXiv* **2020**, arXiv:2004.11886.
33. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
34. Jiao, J.; Tang, Y.M.; Lin, K.Y.; Gao, Y.; Ma, J.; Wang, Y.; Zheng, W.S. Dilateformer: Multi-scale dilated transformer for visual recognition. *IEEE Trans. Multimed.* **2023**, *25*, 8906–8919. [CrossRef]

35. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
36. Ke, J.; Lu, Y.; Shen, Y.; Zhu, J.; Zhou, Y.; Huang, J.; Yao, J.; Liang, X.; Guo, Y.; Wei, Z.; et al. ClusterSeg: A crowd cluster pinpointed nucleus segmentation framework with cross-modality datasets. *Med Image Anal.* **2023**, *85*, 102758. [CrossRef]
37. Wang, S.; Song, J.; Lien, J.; Poupyrev, I.; Hilliges, O. Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology, Tokyo, Japan, 16–19 October 2016; pp. 851–860.
38. Berenguer, A.D.; Oveneke, M.C.; Alioscha-Perez, M.; Bourdoux, A.; Sahli, H. Gesturevlad: Combining unsupervised features representation and spatio-temporal aggregation for doppler-radar gesture recognition. *IEEE Access* **2019**, *7*, 137122–137135. [CrossRef]
39. Ahmed, S.; Kim, W.; Park, J.; Cho, S.H. Radar-Based Air-Writing Gesture Recognition Using a Novel Multistream CNN Approach. *IEEE Internet Things J.* **2022**, *9*, 23869–23880. [CrossRef]
40. Ahmed, S.; Cho, S.H. Hand gesture recognition using an IR-UWB radar with an inception module-based classifier. *Sensors* **2020**, *20*, 564. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.