Img2Spec: A Generalized Cross-modal Gesture Recognition Framework Based on Local Descriptors

Yongpan Zou, *Member, IEEE*, Weiyu Chen, Yunting He, Feihang Dong, Victor C. M. Leung, *Life Fellow, IEEE*, and Kaishun Wu, *Fellow, IEEE*

***Abstract*—Wireless sensing-based human gesture recognition (HGR) is widely used in smart homes due to its contactless operation and privacy protection. However, deploying recognition systems for new tasks typically requires labor-intensive data collection. Existing methods leverage data from different modalities to construct cross-modal recognition frameworks, which reduce reliance on target modality data. Nonetheless, these methods usually require strong category consistency between source and target modalities to reduce domain differences, which limits their generalization across multiple modalities. To eliminate the need for category consistency and achieve across multi-modal recognition, we propose *Img2Spec*, a generalizable cross-modal gesture recognition framework, which pretrains on image datasets and effectively recognizes gestures captured by wireless signals using minimal target modality data. We propose the modal-aware attention local matching (MAALM) module, which combines mutual nearest local descriptors and attention mechanisms to enhance inter-class separability and reduces intra-class variation, enabling efficient knowledge transfer to various sensing modalities. Experiments show that with five support samples, *Img2Spec* achieves recognition accuracies of 95.47%, 93.49%, and 88.44% for acoustic, WiFi, and millimeter wave gestures respectively. With one support sample, our method achieves an accuracy improvement of up to 15.40% over the baseline. Our framework totally eliminates the need for training data and supports gesture recognition with different sensing modalities.**

***Index Terms*—Cross-modal Framework; Gesture Recognition; Local Descriptors**

I. Introduction

Received 30 December 2024; revised 29 June 2025; accepted 28 July 2025. This work was supported in part by China NSFC under Grant 62172286 and Grant U2001207, in part by Key Project of Department of Education of Guangdong Province under Grant 2024ZDZX1011, in part by Youth S& T Talent Support Programme of GDSTA under Grant SKXRC2025231, in part by Shenzhen Science and Technology Program under Grant JCYJ20250604181429038, in part by the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things under Grant 2023B1212010007, in part by 111 Center under grant D25008, in part by Shenzhen Science and Technology Foundation under grant ZDSYS20190902092853047. *(Corresponding author: Kaishun Wu.)*

Yongpan Zou, Weiyu Chen, Yunting He, and Feihang Dong are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (E-mail: yongpan@szu.edu.cn; richard-cwy426@gmail.com; ytingrec@gmail.com; dfh1187933105@gmail.com).

Victor C. M. Leung is with the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, China (E-mail: vleung@ieee.org).

Kaishun Wu is with the Information Hub, Hong Kong University of Science and Technology, Guangzhou 511453, China (E-mail: wuks@ust.hk).

Manuscript received March xx, 2024

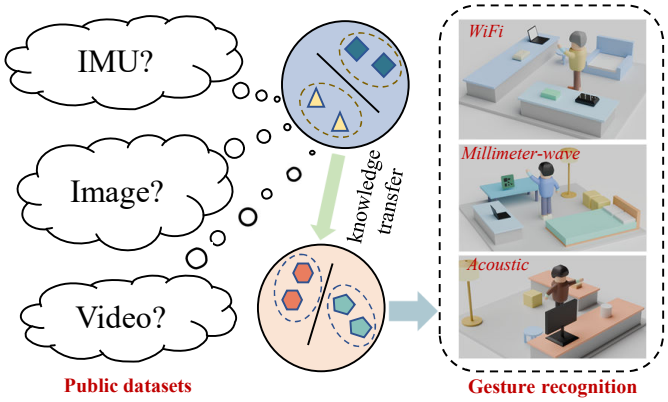


Fig. 1. The sketch of a cross-modal framework leveraging knowledge from public datasets for gesture recognition

WITH the rise of smart devices, wireless sensing-based gesture recognition has become a popular contactless method for human-computer interaction. Many gesture recognition methods have been developed in recent years. These include approaches based on acoustic [1]–[5], WiFi [6]–[10], millimeter wave [11]–[15], inertial measurement units (IMU) [16], [17], and vision-based methods [18], [19]. IMU-based methods often require users to wear devices, which can be uncomfortable. Meanwhile, vision-based methods rely on cameras, which raise privacy concerns. In contrast, wireless sensing-based gesture recognition does not require contact, offers high comfort, and protects privacy. These advantages make it a research focus in the field of human-computer interaction.

However, current wireless sensing-based gesture recognition methods heavily rely on large amounts of training data, which must be collected separately for each modality and specific task. This process is time-consuming and labor-intensive. For example, SonicASL [4] requires 20 acoustic samples per gesture class, the research [8] needs 100 WiFi samples per sentence, and the work [11] collects over 30 millimeter wave radar samples. To address these challenges, researchers have proposed several solutions: (1) Generating target modality data from source modality data [20]–[25]; (2) Leveraging source modality data to assist model training [26]–[28]. While these methods reduce data collection costs and improve performance, they still have several limitations. First, methods that generate target-modality data from source modality data typically require highly similar or consistent activity categories to minimize the discrepancy between the source and target

modalities. This constraint is called category consistency, which restricts the task categories and cannot be extended to more classes. Second, these solutions, designed for a single target modality, often lack adaptability and fail to generalize across new modalities and diverse environments. Third, when video data is used as the source modality, complex techniques are needed to handle occlusion. As a result, this naturally raises an interesting question: *can we design a generalizable cross-modal gesture recognition framework, as illustrated in Fig. 1, that transfers knowledge from a single source modality to multiple target modalities without the category consistency constraint?*

We propose *Img2Spec*, which learns from unrelated open-source image datasets and applies to recognition tasks across diverse wireless sensing modalities. To achieve this, *Img2Spec* must overcome several key challenges: (1) *How to select an appropriate source modality dataset?* With many public datasets available across modalities like images, videos, and IMUs, choosing a source modality that supports effective knowledge transfer to multiple target modalities is critical. if the source and target modalities do not share common structural patterns, knowledge transfer may fail, and leveraging the source modality for auxiliary training can degrade performance. The selected dataset must support balanced transfer across multiple target modalities to ensure broad applicability. (2) *How to preserve the knowledge learned from the source modality maximally?* Since our framework targets multiple modalities, it is difficult for the categories in the source modality dataset to match all target modality categories. This category inconsistency causes significant differences between the source and target domains, making knowledge transfer difficult. Therefore, we need to design a method to remove category constraints, enabling knowledge transfer across different target modalities while retaining each modality's internal pattern variations to improve generalization. (3) *How can the framework be enabled to adapt across multiple target sensing modalities?* Differences in data representation across modalities increase the demand for flexible feature extractors. In addition, the training strategy must allow the model to leverage knowledge from the source without introducing bias toward any specific target modality.

First, we use three widely adopted wireless sensing signals, namely acoustic, WiFi, and millimeter wave, as the target modalities. These modalities differ in raw data formats: Acoustic data is stored as audio files; WiFi data is represented as CSI channel sequences; Millimeter wave data is stored as point cloud representations. Despite these differences, prior works [2]–[6], [9]–[11], [13], [20], [21], [26]–[31] transform these signals into 2D spectrograms via modality-specific preprocessing, allowing models to exploit visual features for recognition. Specifically, acoustic gestures are transformed to spectrograms using the Short-Time Fourier Transform (STFT) [2], WiFi gestures are represented as doppler frequency shifts (DFS) from CSI channels [10], and millimeter wave gestures are extracted as concentrated position-doppler profiles (CPDP) from point cloud data [11]. Despite the differences in raw signal formats, prior work commonly transforms wireless sensing data into spectrograms, enabling models to leverage visual representations for recognition. This observation suggests that different sensing modalities can be unified under an image-like representation, motivating us to explore image datasets as a general source modality. By treating spectrogram as a bridge, we aim to facilitate knowledge transfer across heterogeneous sensing modalities while reducing the impact of modality gaps.

Fig. 2. The acoustic spectrograms of writing '0'

However, although different target sensing modalities can be correlated with the image modality, significant feature discrepancies still remain. As shown in Fig. 2, motion-induced frequency shifts form distinctive local patterns in spectrograms, which are critical for accurate gesture recognition. In computer vision, models often rely on salient local structures and fine-grained details for classification. This motivates us to investigate whether locality-aware representations can be effectively transferred to wireless sensing modalities. Inspired by recent cross-domain studies [32]–[34], we observe that global feature representations may introduce bias when handling multiple modalities, as they tend to incorporate irrelevant background information. In contrast, local features focus on discriminative regions and are more robust to cross-modal discrepancies. Therefore, we propose to leverage local descriptors to capture discriminative patterns and facilitate effective knowledge transfer. Based on this intuition, we design a modality-aware attention local matching (MAALM) module to enhance cross-modal alignment and improve recognition performance. In addition, to learn general and transferable representations, we adopt unsupervised contrastive learning as the pretraining strategy, followed by modality-specific fine-tuning for downstream tasks.

Based on this design, we propose *Img2Spec*, a generalizable cross-modal gesture recognition framework. It uses unlabeled image dataset for pretraining and requires only a small amount of data for fine-tuning downstream modalities. This approach enables the construction of a high-performance gesture recognition system. *Img2Spec* is not merely a stacking of target modalities. Instead, we propose MAALM to eliminate category consistency between the source and target modalities, demonstrating that local descriptors can preserve source domain knowledge in cross-modal tasks. In summary, the main contributions of this work are as follows:

- We propose *Img2Spec*, a generalizable framework for cross-modal gesture recognition that greatly reduces data collection costs and achieves high performance with just a few samples from target modality. To the best of our knowledge, this is the first work to effectively transfer knowledge from image data to acoustic, Wi-Fi, and mmWave data simultaneously.
- We design the modal-aware attention local matching (MAALM) module, which can effectively reduce back-

ground interference and enhance knowledge transfer by leveraging local descriptors and attention mechanisms.
- Experimental results show that with five support samples, *Img2Spec* achieves accuracies of 95.47%, 93.49%, and 88.44% on acoustic, WiFi, and mmWave gesture recognition tasks, exceeding the baseline by 5.15%, 4.96%, and 5.78%, respectively. The source codes and datasets will be released upon the acceptance of this manuscript[1].

The rest of this paper is organized as follows. Sec. II discusses related work. Sec. III-B presents problem definition. Sec. IV and Sec. V describe the system design and implementation details of *Img2Spec*. Sec. VI presents the evaluation results. Sec. VII discusses some related issues. Sec. VIII concludes our work.

II. RELATED WORK

A. Framework for Cross-modal Recognition Tasks

To reduce data collection costs, existing studies investigate cross-modal solutions. Some approaches [20]–[25] generate target modality data for training using data from alternative modalities. For instance, both Vid2Doppler [20] and Midas [21] generate millimeter wave Doppler data from video data. RF Genesis [22] generates millimeter wave data from textual descriptions based on physical simulations, while IMUTube [23] and SignRing [24] generate accelerometer data from videos. mmCLIP [25] enables zero-shot recognition of unseen activities sensed by millimeter wave, using textual descriptions and synthesized RF signals. In addition, other studies [26]–[28], [35] directly leverage data from different modalities to assist in training models for the target modality. IMU2Doppler [26] and RF-CM [27] train millimeter wave-based activity recognition models using IMU and WiFi datasets, respectively. Img2Acoustic [28] implements acoustic gesture recognition by pretraining model on image dataset. However, these methods typically require category consistency between source and target modalities or restrict knowledge transfer to a single modality. As a result, they limit support for multiple target modalities and necessitate careful selection of datasets due to category alignment constraints.

B. Cross-domain Few-shot Learning

Current few-shot learning methods are generally categorized into meta-learning approaches [36], [37] and metric-based methods [38]–[40]. In meta-learning, a representative approach is model-agnostic meta-learning (MAML) [36], which pretrains a model on base tasks and subsequently tests it on new tasks. Metric-based learning methods map data into a feature space and perform classification based on similarity metrics. For example, the prototype network [40] calculates prototype vectors and classifies based on their similarity. Additionally, matching network [38] and relation network [39] employ neural networks to assess the similarity between samples. The query-centric distance modulator (QCDM) [41] generates channel attention weights between samples.

[1]https://github.com/ISAC-GROUP/Img2Spec

Cross-domain few-shot learning (CD-FSL) methods [32]–[34], [42]–[45] focus on addressing challenges related to cross-domain and cross-category learning. For instance, domain-adversarial prototypical network (DAPN) [42] aligns global distributions, while DN4 [32] uses nearest-neighbor deep descriptors for similarity matching, reducing background noise interference. DMN4 [33] utilizes the mutual nearest neighbor relationships between samples to enhance similarity. Existing works focus on single-image modality recognition across different domains, and we innovatively combine CD-FSL with gesture recognition across multiple wireless sensing modalities.

C. Wireless-based Gesture Recognition

Wireless gesture recognition is widely adopted due to its advantages in security and privacy. For acoustic-based gesture recognition [1]–[5], EchoWrite 2.0 [2] recognizes letters and digits using Doppler time-frequency maps. In WiFi-based recognition [6]–[10], Widar3.0 [10] introduces the body-coordinate velocity profile (BVP) as a novel feature to enhance cross-domain adaptability. For millimeter wave-based recognition, methods [11]–[15] such as mHomeGes [11] utilize point cloud features to generate a concentrated position-doppler profile (CPDP) that effectively characterizes the unique features of different arm joints.

The characteristics of these wireless sensing modalities are diverse, and the required sensing modality may vary depending on the environment and recognition accuracy. However, existing methods are typically tailored to specific modalities and lack generalizability across others. Additionally, these methods often rely on large amounts of training data to generalize the model, which limits the rapid deployment of the system. This reveals a major limitation in current research: the absence of a generalizable, low-cost solution for multiple sensing modalities.

III. PRELIMINARY

A. Understanding Cross-modal Knowledge Transfer

Although generic image datasets and wireless modalities differ in semantic meaning and physical interpretation, effective knowledge transfer can still be achieved between them. In this paper, knowledge transfer does not refer to high-level semantic information, but to the reuse of structural representations captured at low and intermediate levels. Pre-trained models learn generic local patterns such as edges, textures, and spatial continuity, which are not unique to generic images. Such structural patterns are also inherently present in spectrogram representations of wireless signals. Due to motion-induced Doppler effects, wireless signals exhibit characteristic frequency shifts over time, forming continuous and structured patterns in the spectrogram domain. These patterns encode meaningful physical information of gestures and are structurally analogous to edges or contours in generic images. As a result, convolutional filters trained on image datasets can be naturally activated by these structurally similar patterns, enabling the reuse of learned feature detectors across modalities.

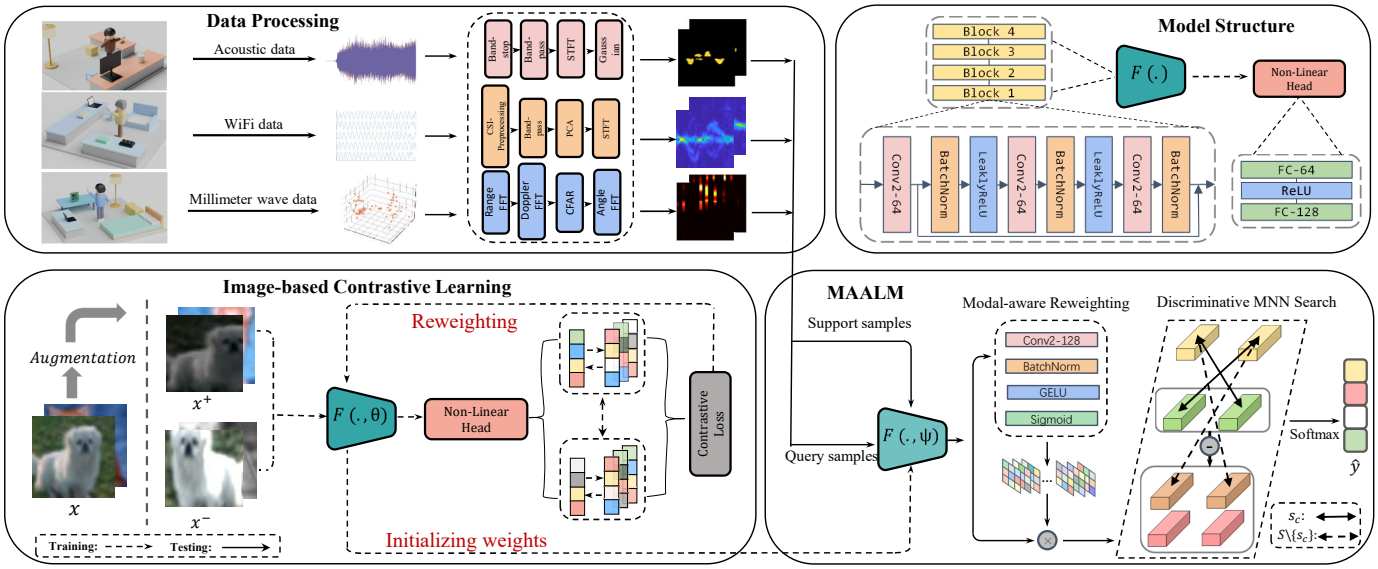


Fig. 3. The system overview of *Img2Spec*

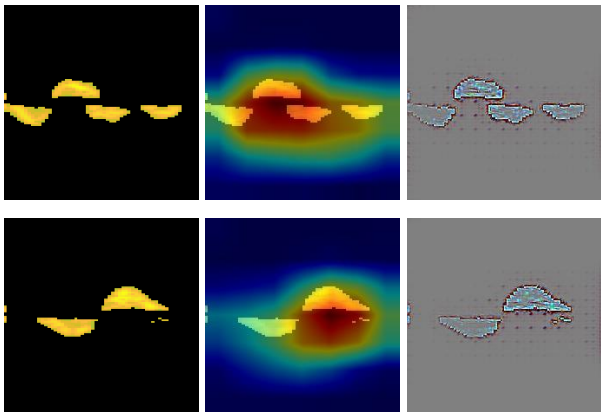


Fig. 4. The visualization of acoustic spectrograms of two different gestures. From left to right: spectrogram, Grad-CAM, and guided Grad-CAM.

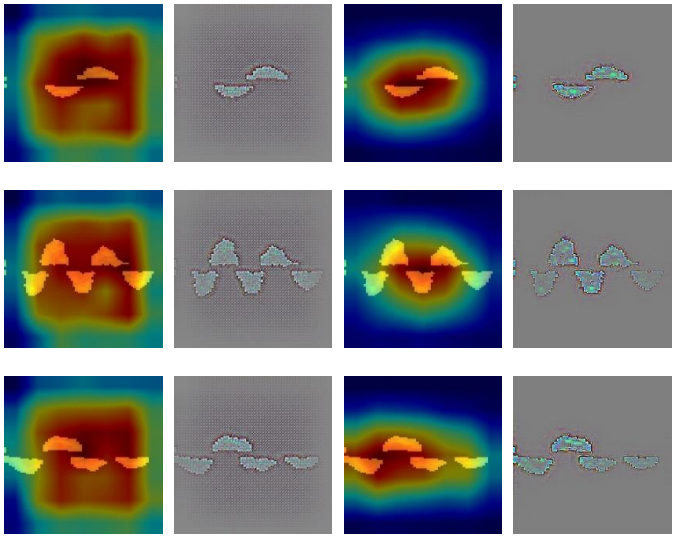


Fig. 5. The visualization of acoustic spectrograms with and without image-based pretraining. From left to right: Grad-CAM (w/o pretraining), guided Grad-CAM (w/o pretraining), Grad-CAM (with pretraining), and guided Grad-CAM (with pretraining).

This behavior is further illustrated by comparing the feature responses of models with and without image-based pretraining using Grad-CAM and guided Grad-CAM [46]. As shown in Fig. 5, the model without pretraining exhibits scattered and noisy activation patterns, failing to consistently focus on discriminative regions. In contrast, the pretrained model produces concentrated and continuous responses that align well with motion-induced frequency variations in the spectrogram. These observations indicate that image-based pretraining enables the model to capture structurally meaningful patterns, rather than responding to arbitrary or noisy regions.

In our framework, cross-modal knowledge transfer is therefore realized at the level of structural representations rather than semantic concepts. Its effectiveness depends on the presence of shared structural regularities between modalities, rather than their semantic similarity. Consequently, not all source modalities are equally suitable for transfer. When the target data lacks clear and consistent structural patterns, the transferred knowledge becomes less effective. In our setting, different wireless signals can be transformed into spectrograms with well-defined structural patterns, which reduces the modality gap and enables effective transfer from image-based pretraining. This observation also motivates the use of local feature representations in our framework, as they are more capable of capturing and matching such structural patterns across modalities.

B. Problem Definition

Img2Spec is a cross-modal gesture recognition framework that utilizes the abundance of image dataset to develop gesture

recognition systems for acoustic, WiFi, and millimeter wave signals. We assume the source modal dataset $\mathcal{D}_{source}$ consists of image samples x_i with corresponding labels y_i, represented as $\mathcal{D}_{source} = \{(x_i, y_i)\}_{i=1}^{n_{source}}$. In the pre-training phase, we train a general feature extractor $F(.,\theta) : \mathcal{X} \rightarrow \mathcal{Z}$, where $\mathcal{X}$ is the input space and $\mathcal{Z}$ is the feature space. During fine-tuning, we use an N-way K-shot approach to adapt the pretrained feature extractor $F(.,\theta)$, where N is the number of target modal classes, and K is the number of support samples per class. The target modal dataset $\mathcal{D}_{target} = \{(x_i, y_i)\}_{i=1}^{n_{target}}$ has a very small number of samples. $F(.,\theta)$ randomly selects a series of N-way K-shot segments from $\mathcal{D}_{target}$ for parameter adjustment. Each training epoch includes multiple tasks $\mathcal{T}$. In each task $\mathcal{T}_i$, the support set $S = (x_{S_i}, y_{S_i})_{i=1}^{N\times K}$ contains K samples per class, and the query set $Q = (x_{Q_i}, y_{Q_i})_{i=1}^{M}$ includes M query samples. In few-shot learning, support samples are crucial for providing labeled examples of each class, allowing the model to learn and generalize from limited data. Within the same task $\mathcal{T}$, $S \cap Q = \emptyset$. After fine-tuning across several tasks, the model produces a modality-specific feature extractor $F(.,\phi)$. In real-world applications, users can choose a target modality and provide K custom gesture samples as the support set S. During testing, the user's input samples are treated as the query sample q and combined with the pre-stored support set S to predict the specific gesture class using our proposed framework.

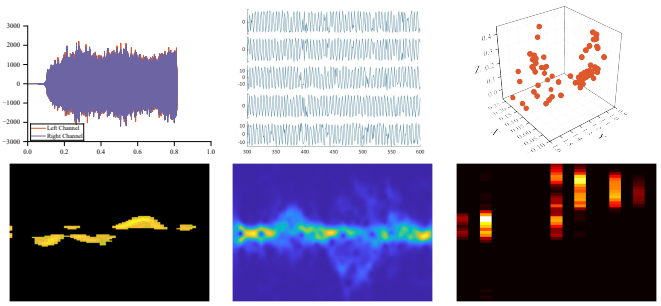
IV. SYSTEM DESIGN

A. System Overview

Img2Spec is a generalizable cross-modal gesture recognition framework pretrained on open-source image datasets and finetuned with a small number of samples from the target modality, enabling high recognition accuracy across various modalities. The framework uses a large set of unlabeled image data for pretraining and fine-tunes with a few labeled samples to ensure the model can adapt to the characteristics of different modalities dynamically. The system workflow is shown in Fig. 3. In the pretraining phase, *Img2Spec* leverages contrastive learning on a large image dataset to train the base model. In the fine-tuning phase, the encoder from the pretraining phase acts as the feature extractor. Users can choose the modality for gesture recognition and provide a small set of labeled samples to fine-tune the model. During testing, signals of different modalities are first converted into spectrograms through modality-specific preprocessing and then passed through the feature extractor. The model transforms the extracted features into local descriptors. The matching module computes classification probabilities based on the relationships among these descriptors. Finally, the model outputs the gesture recognition results.

B. Data Pre-processing

According to the discussion in Sec. I, data from different target modalities can be transformed into image-like representations through appropriate processing. To ensure consistent feature representation across modalities, we convert the acoustic signals, Wi-Fi CSI, and mmWave point clouds into a unified format using modality-specific preprocessing methods as illustrated in Fig. 6.



(a) Acoustic (b) WiFi (c) mmWave

Fig. 6. The feature maps generated from the raw data of acoustic, WiFi, and mmWave modalities

1) Acoustic data: According to the Doppler effect, when a user performs hand gestures within the acoustic field of a speaker, the frequency of the echo signal received by the microphone shifts. The magnitude of this frequency shift, denoted as f_D, depends on the relative movement speed between the finger and the speaker-microphone pair. It is defined as follows:

$$f_D = f_0 \cdot \left| 1 - \frac{v_s \pm v_f}{v_s \mp v_f} \right| \tag{1}$$

where f_0 represents the frequency of the transmitted signal, v_s is the propagation speed of the signal in air, and v_f is the relative motion speed of the finger with respect to the device.

In our experiment, to avoid low-frequency noise, we set the frequency of the signal transmitted by the speaker to 19 kHz. Given that the maximum finger movement speed is approximately 1.5 m/s, the maximum Doppler shift is about 168 Hz, as shown in Eq.(1). Therefore, we focus on the frequency band $[18800, 19200]$ Hz. To isolate the target frequency band and remove static environmental components, we first apply a third-order Butterworth notch filter to isolate $[18985, 19015]$ Hz, followed by a bandpass filter for the desired frequency range. The audio signal is then converted into a spectrogram using Short-Time Fourier Transform (STFT) with a Hamming window, 8192 window size, and 1024 overlap size. After obtaining the spectrogram, we normalize it, smooth it with a Gaussian filter, and apply thresholding.

2) WiFi data: Channel State Information (CSI) describes the propagation of wireless signals between a transmitter and receiver. Wireless signals typically travel through multiple paths. For a subcarrier with N propagation paths, the CSI value between a pair of transmitting and receiving antennas as a function of frequency f and time t is given by:

$$H(f,t) = \sum_{n=1}^{N} A_n(t)\, \mathrm{e}^{-j2\pi f \tau_n(t)} \tag{2}$$

where $A_n(t)$ and $\tau_n(t)$ represent the attenuation factor and propagation delay of the n-th channel, respectively.

According to the Doppler effect, human movement induces phase variations in wireless signals, with the Doppler fre-

quency shift denoted as f_D. The changes in wireless signals caused by gestures can be mathematically represented as:

$$H(f,t) = H_s(f) + \sum_{n \in P} A_n(t) e^{-j2\pi f_D \tau_n(t)} \tag{3}$$

where $H_s(f)$ represents the sum of static path responses (*i.e.* $f_D = 0$) that remain constant over time, and P denotes the set of dynamic paths (*i.e.* $f_D \neq 0$). To remove static environmental components $H_s(f)$ and high-frequency noise, we use a bandpass Butterworth filter to retain the frequency band $[2, 60]$ Hz. Then, we apply a STFT with a Gaussian window on the CSI data, setting the window size to 250 and a 75% overlap. The DFS features capture intensity information across specific time and frequency ranges, offering insight into the user's dynamic state.

3) Millimeter wave data: mHomeGes [11] uses position and velocity information from millimeter wave radar to propose the CPDP. Specifically, the millimeter wave radar captures the user's motion to produce a point cloud in each frame. Each point can be represented as a vector $\vec{p_i} = (x_i, y_i, z_i, d_i, \epsilon_i)$, where i denotes the current frame index. Here, x_i, y_i, and z_i represent the position of the point, d_i represents the point's velocity relative to the radar, and ϵ_i indicates the reflection intensity. A point cloud in the j-th frame is denoted as $P_j = \{\vec{p}i | i = 1, 2, \ldots, n_j\}$, where n_j is the number of points in that frame. Therefore, we can assume that a gesture sequence consists of N frames, it can be represented as $Ges = \langle P_1, P_2, \ldots, P_N \rangle$. CPDP features aggregate reflection intensities at specific positions and velocities, mapping each gesture onto a two-dimensional matrix. It captures the energy distribution of the gesture at each position and velocity. The computed distance and velocity values are normalized to the range $[0, 1]$ and mapped to matrix indices. Formally, CPDP can be represented as:

$$\Pi = [\sum \epsilon_{m,n}]_{I_s \times I_d} \tag{4}$$

where $\sum \epsilon_{m,n}$ denotes the intensity of $\vec{p}$, I_s is the distance index, and I_d is the Doppler radial velocity index. By constructing CPDP, raw point cloud signals are transformed into spectrograms that represent different gestures.

C. Pre-training Phase

In this section, we introduce the design of the feature extractor and the details of pretraining.

1) Design of feature extractor: Img2Spec is a framework designed for gesture recognition across various target modalities. It emphasizes the need for a robust and generalizable feature extractor. However, the distributions, scales, and relational structures of features vary across modalities, which may influence fine-tuning performance. Choosing an appropriate network backbone based on the characteristics of different modalities is essential in practical applications. Overly simple or complex models may fail to adapt to the target modality. Smaller-scale models work better for modalities with uniform data distributions, while larger-scale models are more effective for diverse and complex feature distributions.

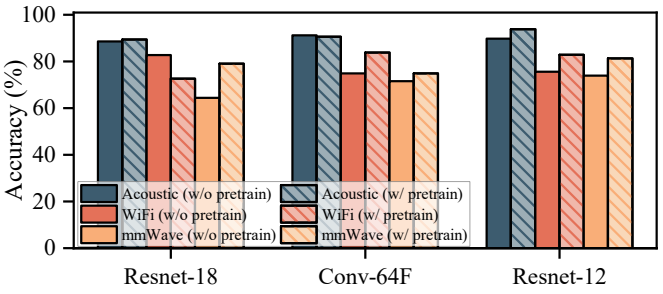


Fig. 7. The impact of pretraining on models of different scales

To address this issue, we compare the performance of models with different scales in pretraining-finetuning and direct-training frameworks. We randomly select one user in each of the three target modalities (acoustic, WiFi, and millimeter wave) for experiments, setting the support sample count to 3. The results are shown in Fig. 7. When the backbone network is ResNet-18 [47], fine-tuning performance on the WiFi modality decreases. For the acoustic modality, a Conv-64F backbone leads to poor performance. In contrast, fine-tuning on the mmWave modality is effective across all three network scales. We attribute this to millimeter wave features being closer to image features, while acoustic and WiFi features differ significantly from images, resulting in the observed performance differences. Considering the differences in data distribution across modalities and the need for a lightweight yet effective architecture, we design the backbone network based on a residual network structure [47]. The network consists of four residual blocks with channel dimensions of 64, 128, 256, and 128 respectively. The proposed model offers advantages in both size and performance, the detailed discussion is provided in Sec. VI-D1. Inspired by [48], we adopt LeakyReLU as the activation function instead of the standard ReLU, which is particularly beneficial for heterogeneous data representations in cross-modal recognition tasks.

2) General Feature Learning via Image Pre-training: Based on the above discussion, we decide to perform preliminary pre-training on open-source image datasets before applying the model to gesture recognition tasks based on different sensing modalities. Given the distinct characteristics of data from different modalities, we aim to learn generalizable representations of 2D features from image datasets during pre-training to maximize its effectiveness. We also expect the pre-training phase to be decoupled from downstream tasks. In other words, the framework should require minimal adjustments regardless of the type or class composition of the image dataset used for pre-training. Contrastive learning is used to capture general features from visual data, reducing the reliance on labeled data. Therefore, during pre-training, we initialize the parameters of the feature extractor using contrastive learning.

Unsupervised contrastive learning uses large amounts of unlabeled data to learn meaningful feature representations [49]. An anchor sample is augmented to create positive pairs, minimizing the distance between similar samples and maximizing the distance between dissimilar ones. We use cosine

similarity in contrastive learning to measure the distance between vectors. Given two vectors $\mathbf{x}$ and $\mathbf{y}$, their distance is computed as follows:

$$Cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}|| \times ||\mathbf{y}||} \tag{5}$$

To ensure that the model focuses on learning general visual representations rather than task-specific classes during pretraining, data augmentation is commonly used to generate diverse samples. In this work, we adopt augmentation strategies including random cropping, scaling, horizontal flipping, color jittering, and random noise injection. Given an input sample x_i, we generate augmented sample pairs (i.e., positive sample pairs) $\langle x_i^-, x_i^+ \rangle$. From a batch of N samples, we generate $2N$ augmented samples. To enhance representation and semantic generalization, we add a nonlinear head after the feature extractor. This module consists of a fully connected layer followed by ReLU activation. The output representations are denoted as $\langle s_i^-, s_i^+ \rangle$.

The construction of positive and negative sample pairs is crucial. From the $2N$ augmented samples, we construct negative pairs. Given a sample s_i, its positive pair is s_i^+, and the remaining samples s_j ($j \neq i$) form the negative pairs for s_i. Therefore, for each s_i, there are $2N$ negative pairs. The contrastive regularization loss is defined as:

$$L_C = -\sum_{i=1}^{N} \log \frac{\exp(Cos(s_i, s_i^+)/\tau)}{\sum_{j \neq i}^{2N} \exp(Cos(s_i, s_j)/\tau)} \tag{6}$$

where τ is the temperature parameter, s_i^+ is the augmented positive sample of s_i, and s_j represents the remaining augmented samples. After pretraining, only the feature extractor is retained.

D. Fine-tuning / Testing Phase

After self-supervised pretraining, we obtain a feature extractor $F(., \theta)$ that can extract features from image samples. Fine-tuning this extractor with a small number of target modality samples (e.g., 1-shot, 3-shot, or 5-shot) adapts the model to specific modalities, resulting in a gesture recognition model $F(., \Psi)$. In this section, we explain why local descriptors are suitable for cross-modal tasks and provide details of our MAALM module.

1) Global and local matching strategies: In few-shot learning, global matching and local matching are common strategies for feature matching. For a query sample, the feature map Z_q is matched with the feature maps Z_{s_i} of support samples from a specific class, where $i = 1, 2, ..., K$. Here, $Z \in \mathbb{R}^{d \times h \times w}$, where d is the dimension of feature channels, and h and w are the height and width of the feature map, respectively.

Global matching strategy: In global matching, the feature of an input sample is compressed into a fixed vector representation. For example, in the prototypical network [40], a prototype vector represents all support samples of a given class. This representation is called the global descriptor. Global descriptors are compared by computing their similarity to determine the class. As shown in Fig. 8, the global descriptor of a query sample, D_q, is obtained by applying average pooling

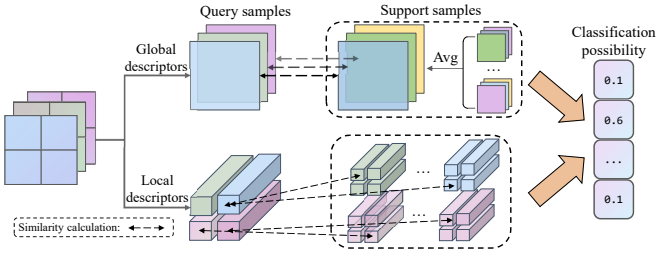


Fig. 8. The illustration of global and local matching strategies

to Z_q, with $D_q \in \mathbb{R}^{d \times 1 \times 1}$. Similarly, the global descriptor for support samples, D_{s_i}, is computed as $D_{s_i} \in \mathbb{R}^{d \times 1 \times 1}$. The prototype vector P is calculated as:

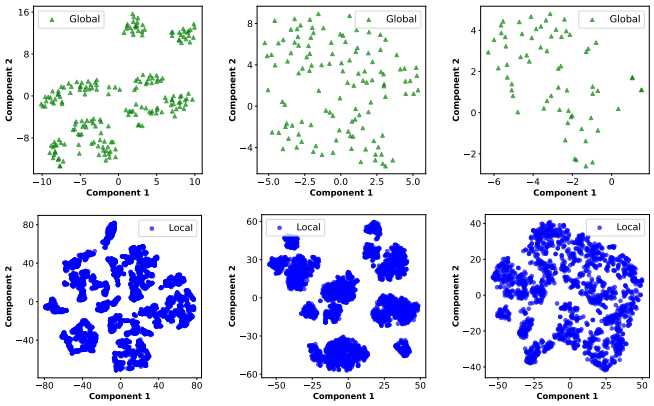
$$P = \frac{1}{K} \sum_{i=1}^{K} D_{s_i} \tag{7}$$

Finally, the query sample's class is determined by the cosine similarity between D_q and P. Global matching is computationally efficient but often loses local feature information. In cross-modal learning, it is particularly sensitive to differences between modalities.

Local matching strategy: In contrast, local matching divides sample features into multiple local regions for comparison. For instance, DN4 [32] uses nearest-neighbor local descriptors to compute similarity for classification. Here, Z_q is represented as local descriptors D_q as shown in Fig. 8, where $D_q \in \mathbb{R}^{h \times w \times d}$, and D_{s_i} for support samples is also in $\mathbb{R}^{h \times w \times d}$. Nearest-neighbor descriptor pairs $\langle D_{q_j}, D_{s_{i_k}} \rangle$ (where $j, k \in h \times w$) are identified, and their cosine similarity is summed to compute the similarity score. This approach focuses on the most valuable information while ignoring common features across classes, such as background or texture. It enhances the domain-specific distinctiveness of the target modality. For example, when support samples are flipped or partially missing, global descriptors may lose information, but nearest-neighbor local descriptors retain their relationships [32]. Local descriptors captures finer sample details, which helps discard irrelevant features and focus only on the important information of the target modality in cross-modal learning.

We conduct experiments to validate the effectiveness of local descriptors in reducing source-target domain discrepancies while preserving intra-domain distinctiveness. In these experiments, we use a pre-trained ResNet-18 [47] as the feature extractor and derive global and local descriptors through average pooling and dimension reordering, respectively. We visualize these descriptors using t-SNE [50]. As shown in Fig. 9, the first row presents the global descriptors of the target modality, while the second row shows the local descriptors. Compared to global descriptors, local descriptors exhibit larger inter-class distances and more distinct clusters. However, the class boundaries of millimeter wave local descriptors remain unclear, suggesting that additional filtering mechanisms can be added to improve class distinction.

In summary, local descriptors often focuses on the most significant regions for classification, this ability enables it to fully use the knowledge learned from source modalities with



(a) Acoustic (b) WiFi (c) mmWave

Fig. 9. The t-SNE visualization of target modals' descriptors, the top represents the global descriptors and the bottom represents the local descriptors.

inconsistent categories. Therefore, local descriptors are ideal for cross-modal tasks.

2) Modal-aware attention local matching: As discussed above, global descriptors tend to suppress fine-grained local cues when there exists a significant domain gap between training and testing data. In cross-modal sensing scenarios, such information loss is further amplified due to modality-specific signal characteristics and heterogeneous feature distributions. Local descriptors preserve spatially localized patterns and exhibit stronger channel-wise correlations, making them more suitable for cross-modal generalization under limited supervision. Moreover, different sensing modalities emphasize distinct local structures, which cannot be uniformly captured by a modality-agnostic matching strategy. Based on these observations, we propose **modal-aware attention local matching** (MAALM), which explicitly adapts local descriptor matching to modality-specific characteristics before similarity computation.

For a given sample X, the feature map $F(X,\Psi) \in \mathbb{R}^{h\times w\times d}$ is viewed as a collection of $r = h \times w$ local descriptors. The descriptor sets for query samples and support samples are denoted as $\mathbf{q}$ and $\mathbf{S} = \bigcup_{c\in C} \mathbf{s_c}$, respectively, where C is the set of all categories, and $\mathbf{s}_c$ represents the descriptor set for category c.

Img2Spec is intended to support multiple target modalities, yet the image-like features derived from different sensing modalities—after the processing described in Sec. IV-B may exhibit substantial variation. Directly applying local matching across modalities risks emphasizing task-irrelevant or modality-specific noise patterns. To address this issue, we introduce a modal-aware attention module prior to local descriptor matching, enabling adaptive emphasis on modality-relevant feature channels. The feature map $F(X,\Psi)$ is passed through an attention module $\mathbf{A}$, as shown in Fig. 3. This module consists of convolutional layers with a kernel size of 3, using the GELU activation function to extract deeper information. A sigmoid function constrains the attention weights within the range $[0,1]$. The modality attention map $W = \mathbf{A}(F(X,\Psi)) \in \mathbb{R}^{h\times w\times d}$ is used to reweight the feature map:

$$F'(X,\Psi) = W \odot F(X,\Psi) \tag{8}$$

This attention-guided reweighting suppresses modality-irrelevant activations while preserving discriminative local structures that are consistently informative across sensing modalities. After attention adaptation, we perform local descriptor matching using a mutual nearest neighbor (MNN) strategy to mitigate the influence of background clutter and ambiguous descriptors. Unlike conventional nearest-neighbor matching that independently associates each query descriptor with support descriptors, MNN enforces a bidirectional consistency constraint in the descriptor space. Specifically, for each $q \in \mathbf{q}$ and category c, we find s, the nearest neighbor of q in $\mathbf{s}_c$, denoted as $s = \mathrm{NN}_{\mathbf{s}c}(q)$. Similarly, we locate $\tilde{q}$, the nearest neighbor of s in $\mathbf{q}$, denoted as $\tilde{q} = \mathrm{NN}_{\mathbf{q}}(s)$. If $q = \tilde{q}$, q and s are considered a mutual nearest neighbor pair. To enhance the discriminative power of these selected mutual nearest neighbor pairs, we incorporate the concept of positive and negative samples, strengthening the relationship with support samples of the same category while weakening connections to samples from other categories. The classification contribution of each mutual nearest neighbor pair is formulated as:

$$\hat{c} = \arg\max_{c\in C} \sum_{q\in\mathbf{q}} \left(Cos(q, \mathrm{NN}_{\mathbf{S}_c}(q)) - Cos(q, \mathrm{NN}_{S\setminus\mathbf{s}_c}(q)) \right) \tag{9}$$

where $\hat{c}$ represents the predicted category, Cos is the cosine similarity function, and $S \setminus \mathbf{s}_c$ denotes the set of descriptors in S excluding those in $\mathbf{s}_c$. This discriminative approach leverages information from the query samples and enhances inter-category distinction.

Based on the local matching strategy, we utilize attention and mutual nearest neighbor mechanisms to help the model retain knowledge learned during pretraining, effectively distinguish between different classes, and dynamically adapt to various sensing modalities.

V. IMPLEMENTATION

A. Datasets

1) Training dataset: In the experiment on overall performance, we generally select **CIFAR-10** [51], a benchmark dataset for general object recognition, consisting of 60,000 color images across 10 classes, 6,000 images per class. Each image has a resolution of 32×32. To validate the generalizability of our framework, we also pretrain the model on different datasets in the subsequent analysis.

ImageNet [52]: ImageNet is a large-scale image database containing over 14 million images across more than 20,000 categories. These images encompass a wide range of common objects and scenes encountered in daily life.

MiniImageNet [38]: As a mini-version of ImageNet [52], this dataset is widely used for few-shot recognition tasks. It contains 100 classes, with 600 images per class. Each image has a resolution of 84×84.

MNIST [53]: The MNIST dataset is a collection of handwritten digits used for image classification tasks. It contains

60,000 training images and 10,000 test images, with 10 classes (digits '0'-'9'), each image has a resolution of 28×28.

Omniglot [54]: Omniglot is a handwritten character dataset, containing 1,623 categories from 50 different alphabets. Each category consists of 20 samples, with image resolution of 105×105. It is commonly used for few-shot learning tasks.

For pre-training, the images are all resized to 224×224.

2) Testing dataset: We evaluate the *Img2Spec* framework across three modalities: acoustic, WiFi, and millimeter wave. For each modality, we select gesture datasets, including a self-collected dataset and two public datasets: Widar3.0 and mHomeGes. Each dataset varies in terms of gesture categories and includes customized gestures. During fine-tuning, all feature maps are resized to 84×84 to ensure consistency in feature extraction.

Self-collected Dataset (Dataset 1): We collect an acoustic gesture dataset using an Android application developed in-house. This app controls a smart device speaker to emit a 19 kHz modulated sine wave, while the microphone samples the echo at 44.1 kHz. The echoes are processed to generate time-frequency gesture feature maps, as detailed in Sec. IV-B1. The study involves 10 participants (6 male, 4 female), aged between 18 and 25 years, each performing 10 repetitions of hand-drawn gestures representing digits '0' to '9' in a controlled laboratory environment. Participants use a tablet and maintain a writing distance of $[0, 5)$ cm, with background noise ranging from $[45, 55]$ dB. The tablet is positioned at a $0°$ angle, and the speaker volume is set to 60 dB using a Samsung Tab S2 device. In subsequent experiments, participants perform gestures under varying conditions, including different distances and angles. In total, over 10,000 samples are collected from the 10 participants across 14 domains.

Widar3.0 [10] (Dataset 2): Widar3.0 is a WiFi-based gesture dataset consisting of $12,000$ samples from 16 participants. It covers over 75 domains and includes more than 20 gesture types, such as digits '0' to '9', as well as gestures like push, clap, and slide. For this study, we select data from four users, with position 5 and orientation 3, the number of gesture categories is six. Unlike Widar3.0's approach of converting CSI to BVP features, we transform the data into DFS features.

mHomeGes [11] (Dataset 3): The mHomeGes dataset is designed for gesture recognition in smart home scenarios, comprising $22,000$ samples from 25 participants. It includes 10 custom arm gestures, each repeated 30 times at anchor points between 1.2 m and 3 m (spaced 0.15 m apart) across seven indoor environments such as bedrooms, study rooms, and living rooms. mHomeGes provides raw point cloud data, which we process into CPDP feature maps as described in Sec. IV-B3. For this study, we select gesture data from seven users for evaluation.

B. Implementation

During pre-training, we set the batch size to 128, the dimension of the nonlinear projection head to 64, and the temperature parameter τ to 0.5 in Eq.(6). The model is trained for 30 epochs using the Adam optimizer with a learning rate of 3×10^{-4}, retaining only the feature extractor. For fine-tuning, users provide K target modality samples as required.

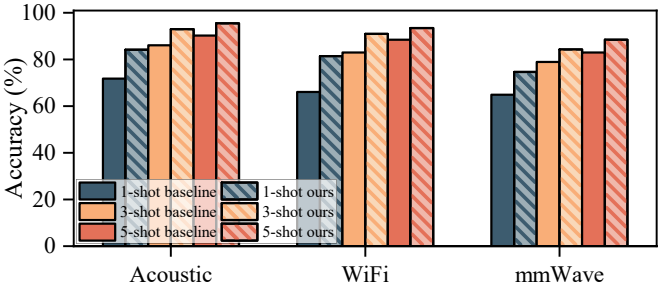


Fig. 10. *Img2Spec*'s overall performance

The model is trained with a batch size of 5 using Adam and a learning rate of 1×10^{-3} for 10 epochs, with each epoch consisting of 10 episodes. To reduce user interaction costs, *Img2Spec* does not rely on validation samples during fine-tuning. Therefore, we retain the model from the final epoch for testing. All data processing and model training are done on a server with an Intel Xeon E5-2686 v4 CPU, an NVIDIA RTX A6000 GPU, and 440 GB of RAM, using the PyTorch library (Python 3.8). Widar3.0's data processing is performed in MATLAB on a PC with an Intel Core i9-10900K CPU.

VI. EVALUATION

A. Evaluation Metric

For performance evaluation, we conduct assessments on three distinct gesture recognition tasks across different modalities, using average accuracy as the evaluation metric. The accuracy is defined as:

$$Accuracy = \frac{N_{correct}}{N_{all}} \tag{10}$$

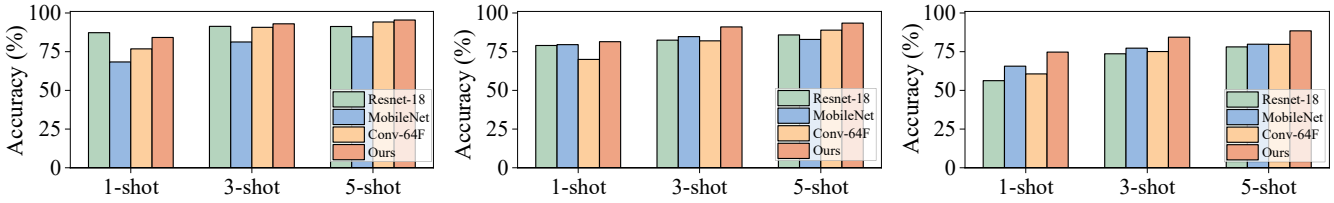
where $N_{correct}$ and N_{all} denote the number of correctly predicted samples and the total number of test samples, respectively. Specifically, we perform fine-tuning and testing using an N-way K-shot approach.

B. Baseline

To demonstrate the effectiveness of *Img2Spec* in enhancing cross-modal hand gesture recognition, we use models trained directly on target modality data as baselines for comparison. These baseline models share the same feature extractor and classification module as *Img2Spec*, ensuring a fair comparison.

C. Overall Performance

We evaluate the performance of *Img2Spec* across different target modalities, including a custom dataset (Dataset 1) and two public available datasets (Datasets 2 and 3). In the N-way K-shot setting, we select K samples randomly from the dataset for fine-tuning, with K samples as support samples and one sample as a query sample. The remaining samples are used only for testing. For acoustic gesture recognition, we focus on the core experiments outlined in Sec. V-A. For WiFi-based gesture recognition, we evaluate data collected in a classroom environment, with coordinates of $(0.91, 0.91)$ m



(a) Acoustic modality (b) WiFi modality (c) mmWave modality

Fig. 11. Performances in different feature extractor for different modalities

TABLE I: Comparison with different classification methods

Method	Acoustic			WiFi			Millimeter wave		
	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
Prototypical network [40]	69.19%	79.34%	80.96%	72.01%	73.82%	76.18%	61.97%	68.18%	72.57%
Relation network [39]	28.49%	48.11%	36.56%	56.45%	59.66%	74.29%	35.74%	39.55%	27.17%
Linear probe	26.36%	34.99%	39.34%	37.74%	33.71%	55.97%	26.95%	28.09%	27.53%
DMN4 [33]	81.62%	92.17%	94.71%	80.91%	89.94%	92.96%	71.72%	82.63%	88.41%
DN4 [32]	82.49%	91.95%	95.05%	80.32%	89.37%	92.49%	70.72%	82.14%	87.81%
MAALM (Ours)	**84.15%**	**93.02%**	**95.47%**	**81.48%**	**90.96%**	**93.49%**	**74.69%**	**84.38%**	**88.44%**

and an orientation angle of $0°$. For mmWave gestures, we use data collected at a distance of 1.2 m.

Fig. 10 shows that for the acoustic modality, *Img2Spec* achieves accuracies of 84.15%, 93.02%, and 95.47% for the 1, 3, and 5-shot settings, respectively, with improvements over the baseline of 12.29%, 6.91%, and 5.15%. As for the Wi-Fi modality, the corresponding accuracies are 81.48%, 90.96%, and 93.49%, with improvements of 15.40%, 7.93%, and 4.96% compared to the baseline. Regarding the mmWave modality, *Img2Spec* achieves accuracies of 74.69%, 84.38%, and 88.44% for 1, 3, and 5-shot settings, outperforming the baseline by 9.74%, 5.47%, and 5.78%, respectively. These results show that *Img2Spec* effectively transfers prior knowledge from the image dataset to various modalities, demonstrating its generalizability and adaptability for multi-modal gesture recognition tasks.

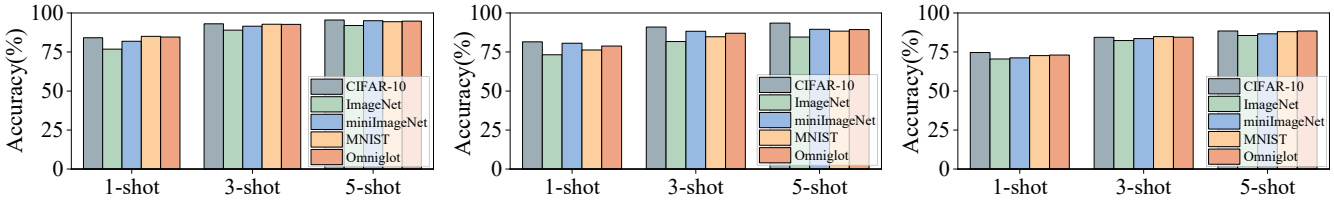
D. Effectiveness Analysis of Framework Components

To evaluate the effectiveness of key components in *Img2Spec*, we replace the feature extractor and classifier with alternatives for assessment in this section.

1) Feature extractor: We design a feature extractor specifically optimized to effectively capture features from spectrograms across different modalities. Additionally, we investigate the recognition performance of various CNN models as feature extractors for different modalities, including ResNet-18 [47], MobileNetv2 [55], and Conv-64F. In this experiment, only the backbone of the feature extractor is modified, while all other settings remain constant. The evaluation results, shown in Fig. 11, clearly indicate that our backbone consistently outperforms the others across all modalities. For the acoustic modality, the 1-shot, 3-shot, and 5-shot settings yield maximum improvements of 15.77%, 11.71%, and 10.83%, respectively. For the WiFi modality, the corresponding improvements are 11.47%, 9%, and 10.57%. In the mmWave modality, the improvements are 18.46%, 10.76%, and 10.30% for the 1-shot, 3-shot, and 5-shot settings, respectively. These results suggest that the performance of the same backbone varies across modalities, likely due to the inherent differences in the data characteristics. The more complex ResNet-18 model may lead to overfitting on simpler data, while the simpler Conv-64F may underperform on more complex data. Furthermore, our backbone has a compact size of only 9.89 MB, which is 1.15 times that of MobileNetv2, making it suitable for deployment on mobile devices. In conclusion, the selected feature extractor is both effective and practical for use in *Img2Spec*.

2) Classification module: To address the challenges associated with multimodal gesture recognition, we propose a classification module, modal-aware attention local matching (MAALM), which effectively leverages knowledge gained from pretraining on image dataset to achieve superior recognition performance across various modalities. To validate the effectiveness of MAALM, we conduct experiments with a fixed feature extractor and compare it against several CD-FSL methods [56], [57], including a simple linear classifier, global descriptor-based approaches (prototype network [40] and relation network [39]), and local descriptor-based methods (DN4 [32] and DMN4 [33]).

The evaluation results, presented in Table I, demonstrate that MAALM consistently achieves the best performance across the three modalities. Specifically, when provided with three support samples, MAALM outperforms the prototype network by 13.68%, 17.14%, and 16.20% across the respective tasks. This highlights the superiority of local descriptors in capturing inter-class differences in multimodal tasks. Notably, when limited support samples are available, classification methods such as the relation network and linear classification head often fail to converge. To facilitate a fair comparison, we adjust the batch sizes to 100 and 16 for fine-tuning these methods. Even under these conditions, MAALM significantly outperforms



(a) Acoustic modality (b) WiFi modality (c) mmWave modality

Fig. 12. Recognition performance of *Img2Spec* pretrained on different source image datasets

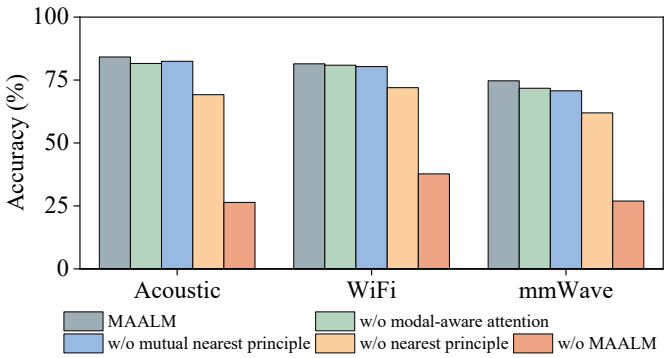


Fig. 13. The results of ablation study

both the relation network and the linear classification head, with an average improvement of 47.52% across all scenarios. Furthermore, compared to DN4 and DMN4, both of which also rely on local descriptors, MAALM achieves maximum improvements of 2.53%, 2.24%, and 1.00% in the 1-shot, 3-shot, and 5-shot settings, respectively. These results validate the effectiveness of the proposed discriminative mutual nearest-neighbor selection mechanism, which enhances inter-class separability and reduces intra-class variance. Additionally, the modal-aware attention mechanism further refines feature representations based on specific tasks, boosting recognition accuracy.

In summary, MAALM consistently delivers superior performance across different modalities and support sample sizes. These findings further demonstrate that the proposed MAALM module effectively leverages prior knowledge to enhance inter-class separability and improve recognition performance in multimodal scenarios, achieving superior results compared with multiple CD-FSL baselines.

E. Generalization Across Diverse Pre-training Datasets

Our goal is to design a universal framework that can be pretrained on open-source image datasets and then applied to various wireless sensing modalities for gesture recognition. Sec. VI-C demonstrates the effectiveness of our cross-modal framework. Sec. VI-D1 verifies that our designed feature extractor can learn generalizable knowledge. Sec. VI-E confirms that our classification module, MAALM, effectively transfers the knowledge learned from image datasets to different wireless sensing modalities, enabling high-accuracy classification. However, this does not yet validate our ultimate objective—a universal cross-modal framework. To verify this, we pretrain the model on different open-source image datasets and then evaluate its performance on gesture recognition tasks across different sensing modalities. This experiment aims to demonstrate that our framework can be pretrained on any image dataset, minimizing user costs and accelerating the deployment of multimodal gesture recognition systems. We conduct pretraining on CIFAR-10 [51], ImageNet [52], miniImageNet [38], MNIST [53], and Omniglot [54], followed by testing.

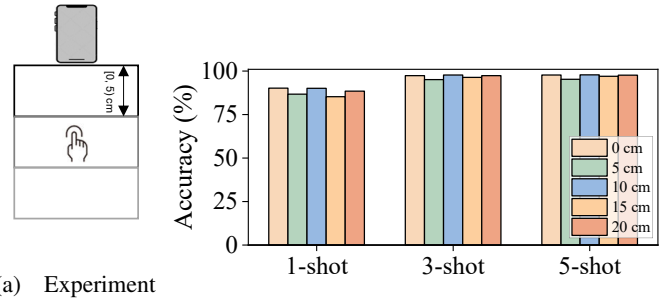
As shown in Fig. 12, the average recognition accuracies for the 1-shot, 3-shot, and 5-shot settings are as follows: For acoustic gesture recognition, the accuracies are 82.48%, 91.80%, and 94.34%. For WiFi gesture recognition, the accuracies are 78.08%, 86.55%, and 89.07%. For millimeter wave gesture recognition, the accuracies are 72.44%, 83.94%, and 87.43%. These results indicate that regardless of the image dataset used for pretraining, *Img2Spec* consistently achieves strong recognition performance across different modalities. Furthermore, the findings highlight the superiority of our proposed feature extractor and classification module.

F. Ablation Study

In this section, we perform an ablation study on MAALM to evaluate the effectiveness of its design. We start with a global feature based linear classifier and progressively add local descriptor matching and subsequent components to quantify their individual contributions. The results are shown in Fig. 13. All three target modalities achieve the highest recognition accuracy when classified using MAALM. Removing either the modal-aware attention or the discriminative mutual nearest neighbor mechanism reduces accuracy, indicating that these components are beneficial when differences exist between target modalities. Classification without local descriptors results in a substantial performance drop, demonstrating that local descriptors are essential for cross-modal recognition. Finally, when the MAALM module is completely removed, the model's classification ability becomes almost unusable. These results confirm that MAALM, as the core component of *Img2Spec*, is effective across different target modalities and facilitates cross-modal gesture recognition.

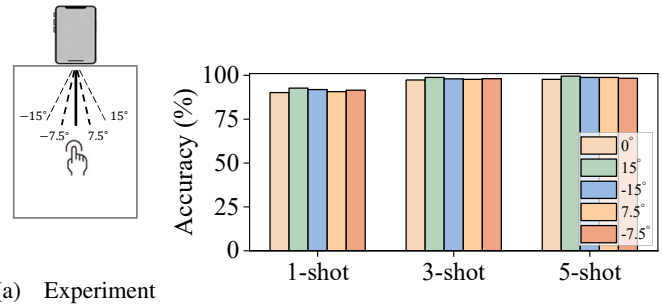
G. Robustness Study

1) Acoustic: Specifically, we assess *Img2Spec* on acoustic tasks across different writing distances, angles, and speaker output power levels. Finally, we involve users from different age groups to further validate the robustness of our approach.



(a) Experiment setup (b) The performance of recognition

Fig. 14. The impact of distance for acoustic modal



(a) Experiment setup (b) The performance of recognition

Fig. 15. The impact of angle for acoustic modal

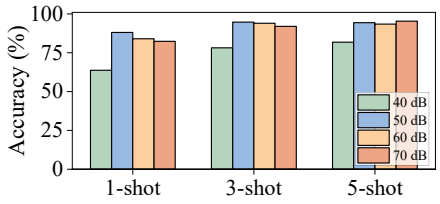


Fig. 16. Acoustic's recognition performance in different volumes

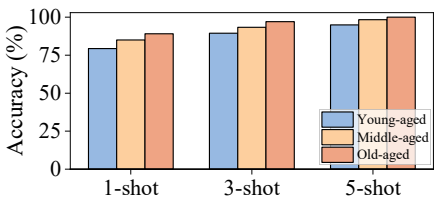


Fig. 17. Acoustic's recognition performance in different age groups

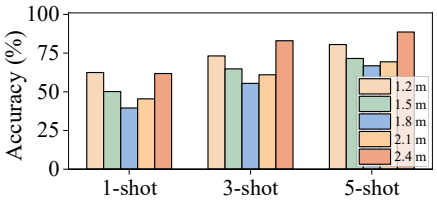


Fig. 18. mmWave's recognition performance in different distances

Distance: In real-world applications, the distance at which users perform gestures can vary. To simulate this variability, participants perform gesture writing at 5 cm intervals, as illustrated in Fig. 14a. The starting distance ranges from $(0, 5]$ cm and extends to $(20, 25]$ cm. The system's performance across these distances is shown in Fig. 14b, where average accuracies of 87.96%, 97.16%, and 97.51% are achieved for 1-shot, 3-shot, and 5-shot scenarios, respectively. However, there is an anomalous behavior observed: recognition accuracy peaks at mid-range distances, while performance degrades at both near and far distances. We attribute this to the fine-grained nature of our gesture set, where even slight perturbations can affect recognition. Since the acoustic signal propagates outward from the speaker as a circular wavefront, the horizontal component of the gesture becomes relatively large at close range and relatively small at long range, leading to performance degradation in both cases.

Angle: In real-world scenarios, users may perform gesture writing at different angles based on their writing habits. To assess the impact of writing angles, we test the system under interaction angles of $0°$, $7.5°$, $-7.5°$, $15°$, and $-15°$, as shown in Fig. 15a. The results present in Fig. 15b, indicating that the system achieves an average accuracy of 91.40% in the 1-shot scenario. When provided with five support samples, recognition accuracy increases significantly to 98.65%. Notably, *Img2Spec* does not achieve its best performance at $0°$, but rather at $15°$. This occurs because the speaker and microphone on the tablet used for data collection are not centrally aligned. Nevertheless, these results demonstrate that the proposed framework exhibits robustness to variations in angle for acoustic gesture recognition.

Volume: Smart devices vary in type, and their speaker output power also differs. To evaluate the recognition performance of *Img2Spec* under different output power levels, users perform gesture writing at four power levels (40-70 dB, with 10 dB intervals). As shown in Fig. 16, recognition accuracy drops significantly at 40 dB, reaching only63.70%, 78.21%, and 81.81% under 1-shot, 3-shot, and 5-shot conditions, respectively. This decline is likely due to the low speaker output power, which weakens the received echo energy, making it difficult for gestures to be effectively represented in spectrograms. When the output power is between 50-70 dB, *Img2Spec* maintains stable gesture recognition performance. Most commercial smart devices can produce sound above 50 dB, mitigating the low performance observed at 40 dB.

User's age: In the evaluation presented in Sec. VI-C, participants are aged between 18 and 25 years. Users from different age groups may exhibit distinct handwriting styles. To assess the impact of age, we evaluate gesture recognition performance across three age groups: young (average age 10), middle-aged (average age 35), and elderly (average age 60). Notably, the model is re-finetuned for this experiment. The results, shown in Fig. 17, indicate that overall gesture recognition accuracy improves with age. The elderly group achieves the highest accuracy, reaching 89.13% with just one support sample. Spectrogram analysis reveals that, compared to the young and middle-aged groups, the elderly participants write more slowly. This slower motion results in more distinct local variations in the spectrogram, enhancing recognition. Nonetheless, *Img2Spec* demonstrates robustness across different age groups in acoustic-based gesture recognition tasks.

2) Millimeter Wave: In the mmWave gesture recognition task, we evaluate system robustness at different distances, specifically 1.2 m, 1.5 m, 1.8 m, 2.1 m, and 2.4 m. Since the mHomeGes dataset [11] contains data from different participants at varying distances, only one participant's data from

TABLE II: Recognition performance of WiFi under different torso locations and face orientations with 5-shot samples

Location	1	2	3	4	5
A	74.94%	87.26%	93.21%	87.26%	97.55%
B	79.27%	90.91%	93.35%	87.58%	96.01%
C	84.63%	95.05%	96.83%	78.67%	91.05%
D	79.28%	92.46%	95.61%	78.32%	91.70%
E	85.26%	96.69%	98.38%	83.66%	94.64%

the $[1.2, 2.4]$ m range is used for testing.

As shown in Fig. 18, with five support samples, the recognition accuracies at 1.2 m and 2.4 m are 72.06% and 77.81%, respectively, while performance at 1.8 m drops to 53.94%. Overall, the recognition accuracy exhibits 'V' pattern, consistent with findings from [11]. We infer that this fluctuation is likely caused by interference introduced during data collection. In summary, the proposed framework achieves recognition accuracy suitable for everyday applications in the mmWave gesture recognition task.

3) WiFi: For WiFi-based gesture recognition, we evaluate the system's performance under different user positions and orientations. We adopt specific parameters from Widar 3.0 [10]. Torso location represents the user's coordinates relative to the wireless transmitter and includes positions A to E. Face orientation denotes the user's facing direction relative to the transmitter and includes orientations 1 to 5. When provided with five support samples, the average recognition performance across all positions and orientations is shown in Table II The model does not achieve the highest accuracy at our designated setting of location E with orientation 3. Instead, the highest recognition accuracy of 98.56% is observed at location A with orientation 2, along with several other settings outperforming our chosen configuration. Nevertheless, we consider location E with orientation 3 to be the most reasonable setting, as it allows users to write gestures naturally while maintaining an appropriate distance and facing the WiFi transmitter. We conduct separate evaluations for position and orientation.

Torso Location: With the transmitter as the origin, locations A to E correspond to (1.365, 0.455), (0.455, 0.455), (0.455, 1.365), (1.365, 1.365), and (0.91, 0.91) in meters. The model is fine-tuned using a small number of samples from each user at location E with orientation 3 and tested on samples from all five orientations. As shown in Fig. 19, gesture recognition at location E achieves the highest average accuracy, reaching 80.52% and 95.20% under the 1-shot and 5-shot settings, respectively. This is because location E is at the center of the sensing region. In contrast, location D exhibits the lowest performance, with 1-shot, 3-shot, and 5-shot accuracies of 72.11%, 86.61%, and 90.79%, respectively. The lower accuracy is attributed to the greater distance from the transmitter, leading to signal attenuation after long-distance reflection. Overall, *Img2Spec* achieves robust recognition performance across different user positions, making it suitable for various application scenarios.

TABLE III: The specifications of mobile devices

Device	CPU clock speed	RAM	Listing year
Huawei Mate40 Pro	3.13 GHz	8 GB	2020
Huawei P50 Pro	3.13 GHz	8 GB	2021
Samsung Galaxy S20	2.84 GHz	12 GB	2020

Face Orientation: Orientations 1 to 5 correspond to user-facing angles of $-90°$, $-45°$, $0°$, $45°$, and $90°$ relative to the transmitter. The model is fine-tuned using a small number of samples from each user at location E with orientation 3 and tested on samples from all five orientations. As shown in Fig. 20, orientation 2 achieves the highest average recognition accuracy, reaching 83.10%, 94.19%, and 96.11% for 1-shot, 3-shot, and 5-shot settings, respectively, outperforming the other four orientations. In contrast, orientation 5 exhibits the lowest accuracy, with only 71.26%, 85.69%, and 90.70% under the same settings. This decline in performance is expected, as orientation 5 introduces the largest deviation relative to the transmitter. However, the superior performance of orientation 2 over orientation 3 is somewhat unexpected. A possible explanation is that a moderate angle between the user and the transmitter allows the model to capture additional motion components, thereby improving recognition accuracy. Overall, when provided with five support samples, *Img2Spec* achieves an average recognition accuracy of 93.63% across all orientations, demonstrating strong robustness.

H. Performance Evaluation in Mobile Device

To evaluate the practical deployment of *Img2Spec* in real-world scenarios, we develop a mobile prototype. In this section, we test *Img2Spec*'s fine-tuning time, response time, memory occupation, and power consumption on Huawei Mate40 Pro, Huawei P50 Pro, and Samsung Galaxy S20. The specifications of these devices are listed in Table III. In our experiments, we assume that the user provides 5-shot samples.

1) Fine-tuning time: Fine-tuning time refers to the duration required to adapt the pretrained general model to a specific modality using support samples. In this experiment, we assume the user provides five support samples for model fine-tuning. As shown in Fig. 21, the acoustic modality requires the longest fine-tuning time, followed by the millimeter wave modality, while the WiFi modality requires the shortest. *Img2Spec* achieves the shortest fine-tuning time on Huawei Mate40 Pro, with 202.52 s, 135.22 s, and 194.68 s for the acoustic, WiFi, and millimeter wave modalities, respectively. All results remain under 4 minutes, which is considered acceptable. Furthermore, as shown in Table III, Mate40 Pro is a smartphone released in 2020, suggesting that future smart devices can perform fine-tuning even faster.

2) Response time: Response time refers to the duration between the user's input of raw signals and the device's output of recognition results. In practical applications, response time is a key factor influencing user experience. To comprehensively evaluate the system's response time, we measure three steps:

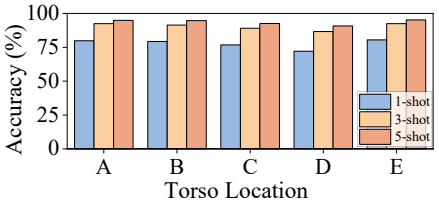


Fig. 19. WiFi's recognition performance in different torso locations

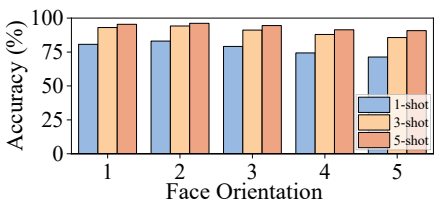


Fig. 20. WiFi's recognition performance in different face orientations

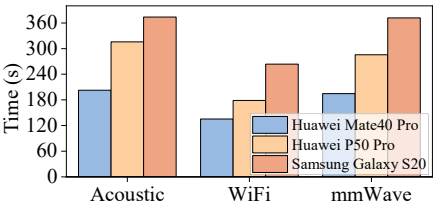


Fig. 21. *Img2Spec*'s fine-tuning time on different mobile devices

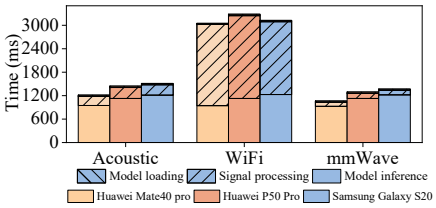


Fig. 22. *Img2Spec*'s response time

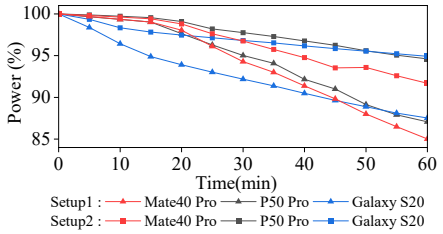


Fig. 23. *Img2Spec*'s battery consumption

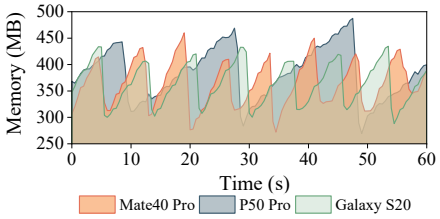


Fig. 24. Memory occupation

TABLE IV: Comparison with existing cross-modal frameworks across multiple perspectives

Framework	Source	Category Consistency	Unseen Class	Multi-modality	Target	Samples	Accuracy
IMU2Doppler [26]	IMU	Yes	No	No	Millimeter	11	76.55%
Vid2Doppler [20]	Video	Yes	No	No	Millimeter	0	81.40%
						54	93.40%
Midas [21]	Video	Yes	No	No	Millimeter	0	84.67%
						27	90.70%
RF-CM [27]	WiFi	Yes	No	No	Millimeter	2	83.19%
						8	96.50%
mmCLIP [25]	Text	Yes	Yes	No	Millimeter	3	76.40%
Img2Acoustic [28]	Image	No	Yes	No	Acoustic	1	82.16%
***Img2Spec* (Ours)**	**Image**	**No**	**Yes**	**Yes**	**Acoustic**	**1**	**84.15%**
					WiFi	**5**	**93.49%**
					Millimeter	**5**	**88.44%**

Note: **Source**: source modality; **Category Consistency**: require category consistency between source and target modalities; **Unseen Class**: support unseen class recognition; **Multi-modality**: support multiple target modalities; **Target**: target modality.

loading the corresponding modality-specific model; performing modality-specific data processing; and inferring the results from the feature maps. As shown in Fig. 22, the model loading time across devices ranges from 30 to 40 ms, and the inference time for different modalities stays within 0.9 to 1.3 s, demonstrating the effectiveness of our lightweight model design. The data processing time varies significantly across modalities. Acoustic data requires slightly more processing time than millimeter wave signals, primarily due to the time-consuming STFT used to convert audio signals into spectrograms. WiFi data processing takes approximately 2 s, since MATLAB code cannot be directly executed on mobile devices. In future work, we aim to migrate WiFi data processing to mobile platforms to reduce processing time. Overall, *Img2Spec* maintains response times within 3.5 s across different sensing modalities.

3) Battery consumption: To evaluate the power consumption of *Img2Spec* on mobile devices, we run the application under two settings on different devices: repeated inference (Setting 1) and idle with the screen on (Setting 2). We record the battery level every 5 minutes over a 60-minute period. As shown in Fig. 23, after 60 minutes of execution, repeated inference increases power consumption by 7.83% on the Samsung Galaxy S20, by 6.66% on the Huawei Mate40 Pro, and by 7.53% on the Huawei P50 Pro compared to the idle setting. These results indicate that *Img2Spec* maintains power consumption within a controllable range across different devices. Meanwhile, the *Img2Spec* prototype system continuously runs for up to 6.67 hours on the P50 Pro. This efficiency benefits from the lightweight feature extractor we design.

4) Memory occupation: To evaluate the memory usage of *Img2Spec* during runtime, we perform continuous inference for 60 s on different devices. As shown in Fig. 24, the maximum memory usage does not exceed 450 MB on the Huawei Mate40 Pro and Samsung Galaxy S20, and remains under 500 MB on the P50 Pro. According to Table III, all these devices have more than 8 GB of RAM. Therefore, we conclude that *Img2Spec* runs smoothly on common commercial smart devices without imposing a significant memory burden.

I. Comparison with Existing Cross-modal Frameworks

In this section, we compare our framework with existing cross-modal approaches from multiple perspectives, such as the requirement of category consistency, the ability to recognize unseen classes, and the support for multiple target modalities. Due to differences in datasets and experimental protocols across prior works, the reported accuracy values are not directly comparable and are included only for reference. Table IV summarizes these comparisons. To reduce the domain gap between the source and target modalities, IMU2Doppler [26], Vid2Doppler [20], Midas [21], and RF-CM [27] require category consistency between the source and target modalities during the auxiliary training process. mmCLIP [25] pretrains the model using text and generated RF signals, followed by fine-tuning with a small amount of real RF signals, leveraging CLIP for unseen class recognition. Although mmCLIP enables the recognition of unseen classes, its average accuracy for 10-class classification is only 76.40%. Img2Acoustic [28], on the other hand, adopts a few-shot learning approach to achieve unseen class recognition, reaching an accuracy of 82.16% with just one sample per class. Overall, these frameworks are designed for specific modalities, such as millimeter wave or acoustic data, which limits their adaptability to other modalities.

In contrast, we design a feature extractor that adapts to the characteristics of multiple modalities and use the fine-tuning strategy and MAALM to enhance the model's generalization across different modalities. For acoustic tasks, the model obtains an accuracy of 84.15% with only one sample per class. For WiFi and millimeter wave tasks, the accuracy reaches 93.49% and 88.44%, respectively, with just five samples per class. In summary, while our framework does not always achieve the highest accuracy under specific settings, it offers several key advantages: (1) it removes the category consistency constraint between the source and target modalities; (2) it supports recognition of unseen classes; and (3) it generalizes across multiple target sensing modalities within a unified framework. These properties highlight the generality and practical applicability of our approach in cross-modal gesture recognition.

VII. DISCUSSION AND FUTURE WORK

A. Generalizability across Target Modalities

Currently, *Img2Spec* is designed for wireless sensing modalities that can convert raw signals into meaningful two-dimensional spectrograms. However, we specifically discuss the transformation of acoustic signals and CSI signals into Doppler time-frequency spectrograms using STFT for acoustic and WiFi-based sensing, and the conversion of captured point cloud signals into CPDP spectrograms for millimeter wave sensing. Many acoustic recognition tasks rely on CIR spectrograms, while numerous WiFi and millimeter wave recognition tasks utilize spectrograms generated through alternative processing methods, highlighting the diversity of data processing techniques. In future work, we plan to further validate the generalizability of the *Img2Spec* framework across a broader range of wireless sensing modality datasets.

B. Expand to More Modalities and Recognition Tasks

In this paper, we evaluate the performance of gesture recognition using acoustic signals, WiFi, and millimeter waves as target modalities. However, this focus does not imply that *Img2Spec* is limited to these three modalities. As a cross-modal gesture recognition framework, *Img2Spec* is adaptable to a broad range of modalities, extending beyond gesture recognition tasks. In future work, we explore the generalization of *Img2Spec* to other target modalities. Additionally, we extend the scope of recognition tasks beyond gesture recognition to applications such as human activity recognition.

C. Reduce Usage Costs

To reduce data collection costs, we employ an unsupervised pretraining approach in combination with a metric-learning-based classification method. At the user level, only a small number of target modality samples are required for fine-tuning. Although only a few samples (1, 3, or 5) are sufficient to build a high-performance gesture recognition system, we still aim to minimize the deployment cost for end users. In future research, we aim to decouple the fine-tuning process from the user level, enabling users to simply select the target modality to achieve effective recognition.

VIII. CONCLUSION

This paper introduces *Img2Spec*, a cross-modal framework for human gesture recognition that achieves strong performance across diverse wireless sensing modalities. It enables users to customize gesture categories and select target modalities, achieving high-performance recognition with minimal fine-tuning overhead. *Img2Spec* achieves modality adaptability through pretraining the feature extractor on public image datasets using unsupervised contrastive learning and integrating the MAALM module based on local descriptors for effective knowledge transfer. Experiments show that the framework leverages general knowledge from images to accurately recognize gestures captured by acoustic, WiFi, and millimeter wave signals. More importantly, *Img2Spec* removes the requirement for category consistency between source and target modalities and demonstrates the effectiveness of local descriptors in cross-modal recognition tasks. We believe this idea will inspire future cross-modal work.

REFERENCES

[1] D. Li, J. Liu, S. I. Lee, and J. Xiong, "Room-scale hand gesture recognition using smart speakers," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 462–475.

[2] Y. Zou, Z. Xiao, S. Hong, Z. Guo, and K. Wu, "Echowrite 2.0: A lightweight zero-shot text-entry system based on acoustics," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 6, pp. 1313–1326, 2022.

[3] K. Ling, H. Dai, Y. Liu, A. X. Liu, W. Wang, and Q. Gu, "Ultragesture: Fine-grained gesture sensing and recognition," *IEEE Transactions on Mobile Computing*, vol. 21, no. 7, pp. 2620–2636, 2020.

[4] Y. Jin, Y. Gao, Y. Zhu, W. Wang, J. Li, S. Choi, Z. Li, J. Chauhan, A. K. Dey, and Z. Jin, "Sonicasl: An acoustic-based sign language gesture recognizer using earphones," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, pp. 1–30, 2021.

[5] W. Chen, C. Zheng, W. He, Y. Zou, and K. Wu, "Ultra write: A lightweight continuous gesture input system with ultrasonic signals on cots devices," in *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2024, pp. 174–183.

[6] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wi-fi," in *Proceedings of the 17th annual international conference on mobile systems, applications, and services*, 2019, pp. 313–325.

[7] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "Signfi: Sign language recognition using wifi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–21, 2018.

[8] L. Zhang, Y. Zhang, and X. Zheng, "Wisign: Ubiquitous american sign language recognition using commercial wi-fi devices," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–24, 2020.

[9] R. Xiao, J. Liu, J. Han, and K. Ren, "Onefi: One-shot recognition for unseen gesture via cots wifi," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, 2021, pp. 206–219.

[10] Y. Zhang, Y. Zheng, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Widar3. 0: Zero-effort cross-domain gesture recognition with wi-fi," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 11, pp. 8671–8688, 2022.

[11] H. Liu, Y. Wang, A. Zhou, H. He, W. Wang, K. Wang, P. Pan, Y. Lu, L. Liu, and H. Ma, "Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 4, no. 4, pp. 1–28, 2020.

[12] D. Salami, R. Hasibi, S. Palipana, P. Popovski, T. Michoel, and S. Sigg, "Tesla-rapture: A lightweight gesture recognition system from mmwave radar sparse point clouds," *IEEE Transactions on Mobile Computing*, vol. 22, no. 8, pp. 4946–4960, 2022.

[13] E. Hayashi, J. Lien, N. Gillian, L. Giusti, D. Weber, J. Yamanaka, L. Bedal, and I. Poupyrev, "Radarnet: Efficient gesture recognition technique utilizing a miniature radar sensor," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.

[14] S. Palipana, D. Salami, L. A. Leiva, and S. Sigg, "Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 5, no. 1, pp. 1–27, 2021.

[15] H. Liu, A. Zhou, Z. Dong, Y. Sun, J. Zhang, L. Liu, H. Ma, J. Liu, and N. Yang, "M-gesture: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3397–3415, 2021.

[16] Q. Zhang, D. Wang, R. Zhao, Y. Yu, and J. Jing, "Write, attend and spell: Streaming end-to-end free-style handwriting recognition using smartwatches," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 3, pp. 1–25, 2021.

[17] Z. Wang, T. Zhao, J. Ma, H. Chen, K. Liu, H. Shao, Q. Wang, and J. Ren, "Hear sign language: A real-time end-to-end sign language recognition system," *IEEE Transactions on Mobile Computing*, vol. 21, no. 7, pp. 2398–2410, 2020.

[18] A. D'Eusanio, S. Pini, G. Borghi, A. Simoni, and R. Vezzani, "Unsupervised detection of dynamic hand gestures from leap motion data," in *International Conference on Image Analysis and Processing*. Springer, 2022, pp. 414–424.

[19] A. D'Eusanio, A. Simoni, S. Pini, G. Borghi, R. Vezzani, and R. Cucchiara, "A transformer-based network for dynamic hand gesture recognition," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 623–632.

[20] K. Ahuja, Y. Jiang, M. Goel, and C. Harrison, "Vid2doppler: Synthesizing doppler radar data from videos for training privacy-preserving activity recognition," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–10.

[21] K. Deng, D. Zhao, Q. Han, Z. Zhang, S. Wang, A. Zhou, and H. Ma, "Midas: Generating mmwave radar data from videos for training pervasive and privacy-preserving human sensing tasks," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 1, pp. 1–26, 2023.

[22] X. Chen and X. Zhang, "Rf genesis: Zero-shot generalization of mmwave sensing through simulation-based data synthesis and generative diffusion models," in *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*, 2023, pp. 28–42.

[23] H. Kwon, C. Tong, H. Haresamudram, Y. Gao, G. D. Abowd, N. D. Lane, and T. Ploetz, "Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–29, 2020.

[24] J. Li, L. Huang, S. Shah, S. J. Jones, Y. Jin, D. Wang, A. Russell, S. Choi, Y. Gao, J. Yuan *et al.*, "Signring: Continuous american sign language recognition using imu rings and virtual imu data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 3, pp. 1–29, 2023.

[25] Q. Cao, H. Xue, T. Liu, X. Wang, H. Wang, X. Zhang, and L. Su, "mmclip: Boosting mmwave-based zero-shot har via signal-text alignment," in *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, 2024, pp. 184–197.

[26] S. Bhalla, M. Goel, and R. Khurana, "Imu2doppler: Cross-modal domain adaptation for doppler-based activity recognition using imu data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–20, 2021.

[27] X. Wang, T. Liu, C. Feng, D. Fang, and X. Chen, "Rf-cm: Cross-modal framework for rf-enabled few-shot human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 1, pp. 1–28, 2023.

[28] Y. Zou, J. Weng, W. Kuang, Y. Jiao, V. C. Leung, and K. Wu, "Img2acoustic: A cross-modal gesture recognition method based on few-shot learning," *IEEE Transactions on Mobile Computing*, 2024.

[29] X. Zheng, K. Yang, J. Xiong, L. Liu, and H. Ma, "Pushing the limits of wifi sensing with low transmission rates," *IEEE Transactions on Mobile Computing*, 2024.

[30] L. Xu, X. Zheng, X. Du, L. Liu, and H. Ma, "Wicamera: Vortex electromagnetic wave-based wifi imaging," *IEEE Transactions on Mobile Computing*, 2024.

[31] X. Leiyang, Z. Xiaolong, and L. Liang, "Vortex em wave-based rotation speed monitoring on commodity wifi," *Chinese Journal of Electronics*, 2024.

[32] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7260–7268.

[33] Y. Liu, T. Zheng, J. Song, D. Cai, and X. He, "Dmn4: Few-shot learning via discriminative mutual nearest neighbor neural network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 2, 2022, pp. 1828–1836.

[34] F. Zhou, P. Wang, L. Zhang, W. Wei, and Y. Zhang, "Revisiting prototypical network for cross domain few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 20 061–20 070.

[35] H. Yang, M. Han, M. Jia, Z. Sun, P. Hu, Y. Zhang, T. Gu, and W. Xu, "Xgait: cross-modal translation via deep generative sensing for rf-based gait recognition," in *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*, 2023, pp. 43–55.

[36] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[37] S. Baik, M. Choi, J. Choi, H. Kim, and K. M. Lee, "Meta-learning with adaptive hyperparameters," *Advances in neural information processing systems*, vol. 33, pp. 20 755–20 765, 2020.

[38] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

[39] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[40] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[41] W. Wu, Y. Shao, C. Gao, J.-H. Xue, and N. Sang, "Query-centric distance modulator for few-shot classification," *Pattern Recognition*, vol. 151, p. 110380, 2024.

[42] A. Zhao, M. Ding, Z. Lu, T. Xiang, Y. Niu, J. Guan, and J.-R. Wen, "Domain-adaptive few-shot learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1390–1399.

[43] Y. Guo, N. C. Codella, L. Karlinsky, J. V. Codella, J. R. Smith, K. Saenko, T. Rosing, and R. Feris, "A broader study of cross-domain few-shot learning," in *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part XXVII 16*. Springer, 2020, pp. 124–141.

[44] P. Li, S. Gong, C. Wang, and Y. Fu, "Ranking distance calibration for cross-domain few-shot learning," in *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, 2022, pp. 9099–9108.
[45] H. Wang and Z.-H. Deng, "Cross-domain few-shot classification via adversarial task augmentation," *arXiv preprint arXiv:2104.14385*, 2021.
[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[48] A. L. Maas, A. Y. Hannun, A. Y. Ng *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1. Atlanta, GA, 2013, p. 3.
[49] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607.
[50] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.
[51] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.
[53] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.
[54] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.
[55] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
[56] Q. Liu, X. Tang, Y. Wang, X. Li, X. Jiang, and W. Li, "Feature transductive distribution optimization for few-shot image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
[57] J. Ren, C. Li, Y. An, W. Zhang, and C. Sun, "Few-shot fine-grained image classification: A comprehensive review," *AI*, vol. 5, no. 1, pp. 405–425, 2024.

Yongpan Zou (Member, IEEE) received the PhD degree from the Hong Kong University of Science and Technology, in 2017. He is currently an associate professor with the College of Computer Science and Software Engineering, Shenzhen University. His research interests include intelligent sensing, mobile computing, and affective computing. He has won several academic awards include IEEE MASS 2014 Best Paper Award, ACM SIGAPP China Rising Star, and etc..

Weiyu Chen received the bachelor's degree from the School of Mathematical Sciences, Shenzhen University, in 2023. He is currently pursuing the master's degree from the College of Computer Science and Software Engineering, Shenzhen University. His research interests cover mobile computing, ubiquitous sensing, and Internet of Things.

Yunting He received the bachelor's degree from the School of Mathematical Sciences, Shenzhen University, in 2024, and is now pursuing the master's degree at the College of Computer Science and Software Engineering, Shenzhen University.Her current research focuses on intelligent sensing and mobile technologies.

Feihang Dong is currently pursuing a bachelor's degree at Computer Science and Software Engineering, Shenzhen University. He will soon join the Zhejiang University to pursue a master's degree, His research interests include spatiotemporal learning, data mining and Internet of Things.

Kaishun Wu (Fellow, IEEE) received the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, in 2011. He is currently a professor in information hub with the Hong Kong University of Science and Technology (Guangzhou). His research interests include wireless communications and mobile computing. He won several best paper awards of international conferences, such as IEEE Globecom 2012 and IEEE MASS 2014.

Victor C. M. Leung (Life Fellow, IEEE) received the PhD degree in electrical engineering from the University of British Columbia, in 1981. He currently is the dean with the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University. His research focuses on wireless networks and mobile systems with more then 60,000 citations. He has received numerous accolades, such as the APEBC Gold Medal, NSERC Postgraduate Scholarships, and IEEE Vancouver Section Centennial Award.