This article has been accepted for publication in IEEE Intelligent Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MIS.2024.3398592

Regular Article

# Unraveling Complexity: An Exploration into the Large-Scale Multi-Modal Signal Processing

Zhenyu Wen, *Zhejiang University of Technology, Hangzhou, China*
Yuheng Ye, *Zhejiang University of Technology, Hangzhou, China*
Jie Su ✉, *Zhejiang University of Technology, Hangzhou, China*
Taotao Li, *Zhejiang University of Technology, Hangzhou, China*
Jinhao Wan, *Zhejiang University of Technology, Hangzhou, China*
Shilian Zheng, *Science and Technology on Communication Information Security Control Laboratory, China*
Zhen Hong, *Zhejiang University of Technology, Hangzhou, China*
Shibo He, *Zhejiang University, Hangzhou, China*
Haoran Duan, *Durham University, Durham, UK*
Yuexiang Li, *Tencent Jarvis Lab, Shenzhen, Guangdong, China*
Yawen Huang ✉, *Tencent Jarvis Lab, Shenzhen, Guangdong, China*
Yefeng Zheng, *Fellow IEEE*, *Tencent Jarvis Lab, Shenzhen, Guangdong, China*

*Abstract—Advanced communication systems and military reconnaissance are increasingly prevalent in high-tech environments, greatly supported by the flourishing in signal processing technologies. The recent exponential proliferation of sensors led to an unprecedented expansion in the scale and diversity of signals across various modalities. Such influx poses significant challenges in effectively integrating multi-modal signal data to deliver comprehensive and interpretive solutions across a diverse range of applications. In this paper, we provide an overview of the core issues, challenges, and future research directions in different stages of developing large-scale multi-modal signal processing models. Additionally, we introduce a prior investigation into signal representation learning, where we propose a contrastive learning-based framework to extract fine-grained signal features under few-shot conditions. Our proposed framework achieves a 24.1% performance improvement over baseline approaches, consistently demonstrating superiority over state-of-the-art methods. The code is accessible in this repository: https://github.com/YYH211/LSM.*

• *Keywords: Multi-Modal Signal Processing, Artificial Intelligence*

## Introduction

Benefiting from the powerful pattern extraction capabilities of deep learning approaches,[1,2,3] the signal processing community starts to discern meaningful patterns from raw signals that may lack visual interpretability. This has led to notable achievements across a range of applications, including cognitive radio, military reconnaissance, threat evaluation, and spectrum monitoring, among others.

Recently, the contemporary digital landscape is witnessing an unprecedented surge in signal generation, stemming from an increasingly diverse array of sensors and devices. The generated signals are often present

Corresponding author: Jie Su and Yawen Huang (e-mail: jieamsu@gmail.com; yawenhuang@tencent.com).

in a multitude of modalities, encompassing a variety of types (e.g., radar signals, WiFi signals, and modulated signals) as well as various formats (e.g., constellation diagrams and spectrograms), driving the need for more sophisticated and efficient methods to manage and interpret this vast influx of data. Nonetheless, existing methods tend to be **task-specific**, **modality-restricted** and **generalization-limited**, making them less capable of handling complex applications and potentially leading to adverse impacts, as illustrated in the following examples.

**Smart Semantic Communication.** The next generation of end-to-end intelligent semantic communication systems equipped with artificial intelligence technology has been proposed due to the problem of transmission error of binary bit streams used in traditional communication. In this new paradigm, the multi-modal information (text, speech, images, etc.) being transmitted is compressed into semantic information at the transmitter and decoded at the receiver. The use of semantic transmission effectively mitigates the bit error rate (BER) problem at low signal-to-noise ratios. However, there is still a huge gap between the current version of semantic communication and its practical application, i.e., the general semantic communication model that has not yet been developed.

As depicted in Figure 1, conventional semantic communication approaches predominantly utilize end-to-end neural networks, rendering them deficient in generalization to novel signals. Consequently, when senders transmit novel signals via conventional semantic communication, they encounter semantic divergence issues, leading to inaccuracies in reconstructing the novel signal message from the receiver 1 perspective. In contrast, on the new paradigm (i.e., multi-modal based smart semantic communication), the model on the receiver 2 side should be able to fuse the multi-modal data knowledge and give an analysis. Such communication paradigm could provide more high quality message transmission ability as well as the generalization ability.

Given these circumstances, the urgency to develop a large-scale multi-modality model for signal processing has become more apparent. The AI community has already seen revolutionary advancements in the vision and language processing fields with the rise of large-scale multi-modality models such as PixelBERT[4] and PaLM-E[5]. However, signals often possess unique attributes that distinguish them from images and texts, presenting new challenges in the development of large-scale multi-modality models. To this end, we comprehensively investigate the primary challenges encountered throughout the lifecycle of designing multi-modality signal models and outline the corresponding key research directions.

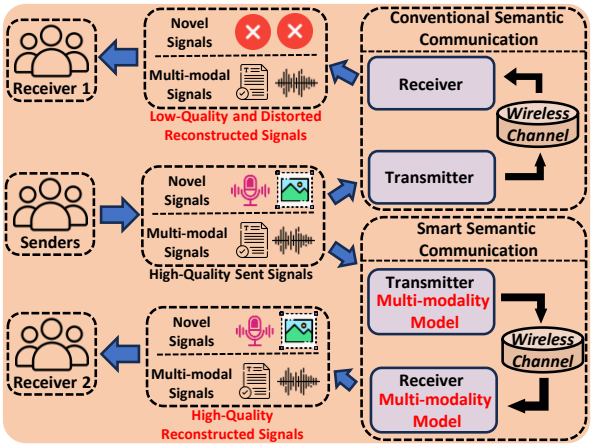


**FIGURE 1.** The difference between the current semantic communication model and the smart semantic communication model.

The main contributions are summarized as follows:

- This is the first position paper that comprehensively investigates the key challenges and corresponding research directions in developing large-scale multi-modal signal models.
- We propose a novel contrastive learning-based framework for learning fine-grained signal representations under few-shot sample conditions, which serves as a foundational step in the design of multi-modality signal model architectures.
- Extensive experiments were conducted, and we studied the proposed framework in detail. The promising results suggested its effectiveness

## Background

Automatic Modulation Recognition contributes as the mainstream task in the communication signal processing domain, which aims to identify the modulation category of the received radio signals. This technology is widely used in spectrum management, interference identification, and electronic reconnaissance systems.

Due to the label scarcity property of communication signals, large-scale signal processing often relies on unsupervised learning approaches, which can be mainly divided into two categories: **Context-based methods**[6,7,8,9], aiming to leverage local-global contrastive optimization to extract meaningful information from raw signals, and **Instance-based methods**[10,11], which aim to leverage local-wise or instance-wise con-

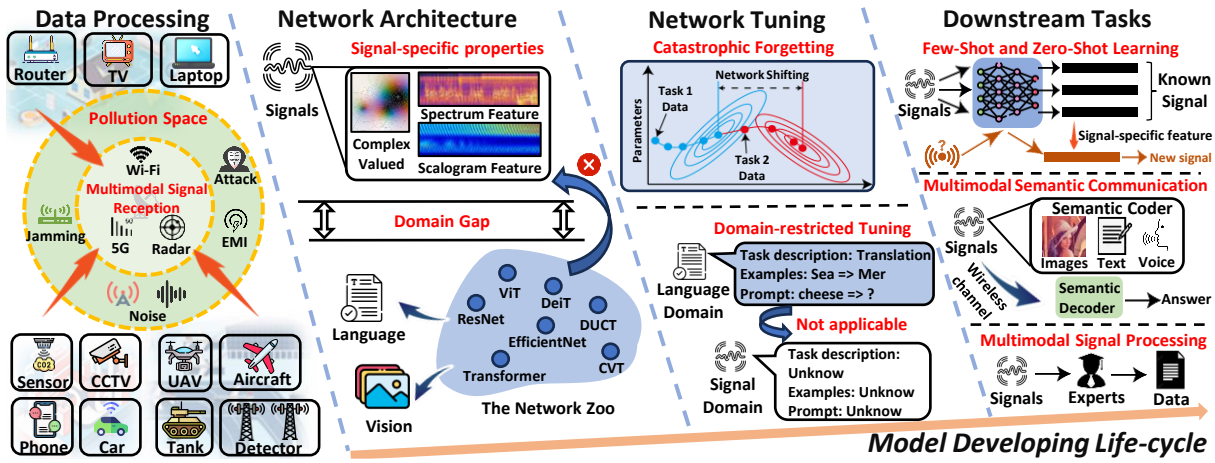


**FIGURE 2.** Challenges on the large-scale multi-modal signal model development life-cycle.

trastive optimization for the same purpose. However, the previous approaches mainly adapted from the vision or language domain neglect the impact of noise on the unique properties (such as frequency, phase, and amplitude) in the signal domain. When applied to data migration between different signal scenarios, these methods may suffer from model collapse and accuracy degradation.

## Challenges & Potential Direction

In this section, we briefly investigate the key challenges during the multi-modal signal model development life-cycle (as presented in Figure 2). Specifically, we analyze the issues encountered throughout the development lifecycle due to the distinct properties of signals, such as noise interference, unique characteristics, and the challenge of objective interpretability.

### Data Processing

**Challenge—Data Contamination.** A problem also studied in the context of large language models,[5] pertains to the possibility of testing content available on the web being unintentionally included in the training data. This can lead to a distortion in the performance evaluation of high-capacity models. The issue of data contamination in signal data (e.g., modulated signals and radar signals) can be markedly more complex, as signals are often transmitted in open environments that are prone to substantial noise interference. Such data contamination could distort the original patterns that exist in signals, resulting in erroneous interpretations and inaccurate model predictions.

**Potential Direction—Noise Reduction.** The problem of model instability due to noise interference in signals is significant, as noise types and sources differ. Traditional noise reduction approaches work well in vision and language processing, but the unique challenges of signal noise due to air transmission require a different solution. One proposal is an adaptive noise filtering module that can adjust to different types of signals. This could include a system combining a conditional signal module and a denoising module, which would apply specific conditions to the noisy signal and clean it at the feature level. This would help the learning system adapt better from the start. Another suggestion is equivariant learning,[12] which maintains the consistency of noisy data under different transformations. This would need a specialized augmentation/transformation function designed around the specific properties of signals, such as symmetry and periodicity, for the optimization of the equivariance constraint. This approach could enhance the denoising capability of the system.

**Challenge—Modality Discrepancy.** Benefiting from the inherent complementary nature and shared semantics across vision-language-speech modalities, cross-modal integration/fusion has been extensively studied and achieved significant advancements in various applications. However, the signal domain presents significant heterogeneity for different modalities, ranging from different categories (e.g., radar and WiFi signals) to varied representations (e.g., constellation diagram and spectrogram). The absence of inherent semantic links largely increases the complexity of feature alignment, which could result in negative information interfering with the future feature extraction

procedure.

**Potential Direction—Multi-Modal Expert Labelling.** The recent advancements in cross-modal/multi-modal learning systems are largely accredited to the existing extensive parallelized cross-modality data (i.e., images with corresponding text descriptions). However, as the aforementioned signal domain limitation, the semantic connection has still not been defined and investigated. A penitential solution for connecting different modalities on a signal domain could be "*Expert Attribute Generation*". For example, similar to the vision fields that describe the observable distinguishing properties (e.g., "black: yes, stripes: yes, eats fish: no") of objects as a text auxiliary information, the observable distinguishing properties of signal domain (e.g., "temporal frequency, amplitude, and pulse") could be introduced. By introducing such intermediate attributes with text and numerical descriptions, the large-scale learning system would be able to map various signals into shared semantic subspace for feature-level cross-modality interaction and information complementariness. Additionally, the designed attributes will largely affect the final performance of the learning system, so expert domain knowledge should be considered during the design.

## Network Architecture

**Challenge—Complex Feature Modeling.** The recent advancements in the vision and language processing community can be largely attributed to the specially designed network architecture (e.g., Transformer and its variants), which is capable of modeling complex patterns, such as spatial-temporal patterns, from images and text tokens. These designs often take into account the inherent properties and characteristics of the image/text data being processed. For example, the attention mechanism is designed to capture the contextual relevance of different image/text pieces. This allows the model to concentrate on the most salient aspects, enhancing its interpretive capabilities and decision-making process. However, signal data is typically present in a complex-valued format, involving both magnitude and phase components. Most deep-learning architectures only utilize half of the spectral input (i.e., the real-valued part), leading to information loss during the feature extraction process. Besides, signal data often exhibits patterns through frequency or time-frequency domain, requiring the network architecture to be designed to capture the interaction between temporal dynamics and complex frequency.

**Potential Direction**—1) **Spectrum-Spatial-Temporal Modelling.** Conventional spatial-temporal modeling is introduced to capture fine-grained dynamic correlated features from a sequence of data samples and achieves decent performance on tasks such as video analysis/summary and time-series modeling. However, signals present an additional dimension (i.e., spectrum) to express patterns or information, which is not able to be captured by the approaches from vision and language domains. A potential solution might be to entangle the designed spectrum learning layer with spatial-temporal modeling. For example, dynamic convolution with kernels of different sizes could be introduced to capture fine-grained spectrum features. Additionally, fusion mechanisms, such as bilinear/multi-stage fusion, on kernel level (e.g., different kernel sizes) and feature level (e.g., spectrum and spatial-temporal features) should be further explored for better feature level interactions and to prevent information loss. 2) **Dynamic Network Capacity.** The large-scale models, with their substantial data processing capabilities, often contain a high capacity for pattern memorization from extensive datasets. The current large-scale models typically possess a pre-defined and fixed capacity, which largely restricts the continual learning ability of the learning system. Such phenomena become more crucial when facing continuous incoming signal data. A potential solution might be to construct an expandable network structure for the designed complex modeling architecture. For example, similar to disentangled representation learning in the vision community, the network architecture could be separated into distinct components[13] (e.g., top, intermediate, and low-level feature extractors) to enable network capacity expansion. The dynamic expansion is triggered when the network capacity reaches the upper boundary so as to provide the foundation for the subsequent tuning procedure. Additionally, the integration of dynamic structure with the aforementioned spectrum-spatial-temporal modeling may be difficult to train, due to the extremely disentangled and complex architecture. Thus, considering specific training paradigms to facilitate dynamic network expansion could also be investigated in the future.

## Network Tuning

**Challenge—Domain-restricted Tuning.** To exceed the performance limits of conventional fine-tuning techniques for downstream tasks, prompt-driven tuning has been well-studied in large-scale language models. The prompt-driven tuning aims to design customized prompts (e.g., different questions) with the correct answers for the tuning process, so as to break through the upper bound. However, restricted by the special

characteristic of the signal domain, where the raw signals often suffer from non-interpretability, the definition of prompt in the signal domain is still unclear.

**Potential Direction**—1) **Network Incremental Learning.** Recently, related incremental learning methods have been introduced into wireless signal recognition, such as the continuous registration of new devices via In-phase and Quadrature (IQ) signals for Internet of Things (IoT) centers, providing incremental update capabilities for aircraft identification systems. However, the similarity of signals from wireless devices under the same communication protocol makes it still difficult for the model to learn the differences between the new classes and the old classes at the same time. Therefore, based on classical incremental learning algorithms, the large model feature extraction capabilities, such as multi-modal signal feature extraction, should be used to provide more fine-grained features to distinguish between new classes and old classes. Providing discriminative multi-modal features in the common space of new and old classes is of significant importance for incremental learning. 2) **Prompt Tuning.** To mitigate the semantic gap and over-fitting problems between downstream tasks and pre-trained models, prompt tuning techniques are currently being extensively researched in natural language processing. However, as mentioned previously, due to the lack of appropriate prompts in the signal domain, they are currently not applied. A potential approach in the signal recognition scenario might be to construct specific conditional templates for different downstream tasks. For example, the conditional auto-encoder can embed conditional variables for downstream tasks, thus continuing to drive the fine-tuning process internally. Another possible way to perform prompt learning on a signal model is to transform the template into signal data, which can be added to a neural network for self-supervised learning. Overall, it is of great research significance to perform prompt tuning on the signal model in order to exploit the potential of the pre-trained model.

## Early Experience

Based on the discussed design philosophy and methodology, we propose a large model framework that can effectively mine knowledge from large-scale signal datasets. As shown in Figure 3, in order to accelerate the deployment for different downstream tasks, a preliminary attempt is made to capture generic signal features for all downstream tasks in a self-supervised learning manner. We explore the two principal directions of noise filtering and spectrum-spatial-temporal modeling. Specifically, we implement a novel generic signal representation learning framework based on contrastive learning. Compared to the traditional contrastive learning-based framework, we make specific modifications to the network structure and data augmentation, and achieve significant improvements in signal recognition.

### Signal Augmentation

To increase the diversity of signal samples so to encourage the feature extraction of the contrastive learning, we simultaneously apply two data augmentation methods (i.e., cropping and rotation). Given an input signal sample, we first crop a sub-segment of fixed length according to a random number which diversifies the samples and focuses on different parts of the original signal to enhance feature extraction. Then, we rotate the sample by a random rotation angle to further enhance the sample diversity.

**Signal Cropping.** For signal cropping, we have designed a patch-based cropping mechanism to preserve local information integrity and reduce the impact of noise on signal recognition. We performed one cropping on each signal to extract a sub-sequence $S'$ of length $I$ from the original data of length $L$, as follows:

$$S' = Crop(L, I, S) = S[\alpha : \alpha + I] \quad for\ I <= L, \tag{1}$$

where $S$ is the original signal data, and $\alpha$ is a random number between 0 and $L - I$. For IQ signals, the modulation information generated should be stable across various short-time delays, indicating that similar modulation information is carried regardless of the cropping location.

**Signal Rotation.** After cropping the signal data, we apply a random rotation to further augment the data features. Considering that a direct rotation of the signal would destroy its integrity, we propose a semantic invariant rotation method. First, we write the signal data in the form as

$$S = x_{real} + x_{imag} \cdot j, \tag{2}$$

where $x_{real}$ and $x_{imag}$ are $1xN$ vectors for the in-phase (I) and quadrature (Q) signals, respectively, and $j$ indicates the imaginary part. From Equation (2), it is obvious that the data distribution of an IQ signal can be represented in the complex plane. For this reason, we retain its complex plane distribution and perform an overall random rotation of it. Therefore, we use the rotation matrix of the two-dimensional plane for IQ to obtain $\hat{S}$:

$$\hat{S} = \begin{bmatrix} \hat{x_{real}} \\ \hat{x_{imag}} \end{bmatrix} = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x_{real} \\ x_{imag}, \end{bmatrix}, \tag{3}$$

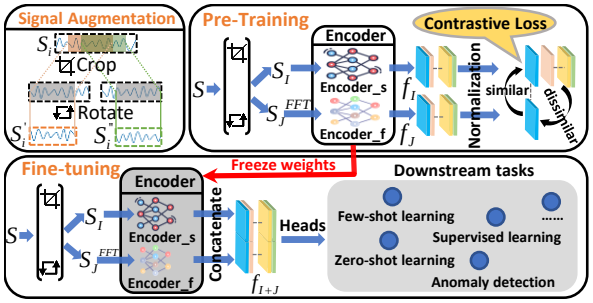


**FIGURE 3.** The framework of the proposed method.

where the $\theta$ is a randomly generated angle ranging from 0 to $2\pi$. Note that the above rotation mechanism will not destroy the original IQ plane distribution.

## Pre-training with Contrastive Learning

Motivated by SimCLR,[10] we design a pre-training framework based on contrastive learning. Contrastive learning learns intra-class and inter-class features by comparing the differences between different augmented samples to enable self-supervised feature extraction. In particular, we explore the spectrum-spatial-temporal modeling in depth to design a special backbone network that can integrate multi-modal signal data.

By exploring the unique representation of signal data, we find that the signal features are mainly expressed in the time and frequency domains. We design two independent encoders to combine the data characteristics of different modalities. (1) We use XCiT[14] as an Encoder_s of the raw time series data since this network mainly uses channel self-attention and can better extract similar semantic features between IQ channels. (2) Encoder_f sub-module acts as a feature extractor for the signal data after the Fast Fourier Transform (FFT), which will help to extract its frequency domain information. The joint extraction of time and frequency domain information will provide a more fine-grained feature representation for downstream tasks of the signal.

After that, we use the NT-Xent[15] loss for our contrastive learning framework, which is defined as follows:

$$\mathcal{L} = -E\left(\sum_{i\in B}\left(log\frac{exp(sim(s_i, s_i^{'})/\tau)}{\sum_{k\in B, k\neq i} exp(sim(s_i, s_k)/\tau)}\right)\right), \quad (4)$$

where $E$ denotes the expectation, $B$ is the current batch size, $s_i$ is the original sample, $s_i^{'}$ is the augmented sample, and $\tau$ refers to the temperature parameter. The cosine similarity is used as the $sim(.)$ function in Equation (4):

$$sim(s_i, s_i^{'}) = \frac{s_i^T s_i^{'}}{\|s_i\| \, \|s_i^{'}\|}, \quad (5)$$

where $\|\cdot\|$ denotes the $l_2$ norm.

## Fine-tuning

In machine learning, the freezing of partial weights is generally used to quickly train new classes. This means that only a small set of the parameters is fine-tuned. We use the same treatment in our fine-tuning phase. As shown in Figure 3, the pre-trained encoder is migrated directly to the fine-tuning phase and is followed by a head for a specific downstream task. As mentioned before, the head (we choose the simplest multi-layer perception as the head) is updated without updating the parameters of the encoder. In detail, we use a small amount of labeled data to fine-tune using the cross-entropy loss. The augmented data can be obtained in two branches after passing through the network, and we concatenate them in the horizontal dimension.

# Initial Results

## Experimental Setup

We evaluate the performance of the proposed model on the commonly used dataset RadioML2016.10a.[16] The dataset contains 11 different signal modulation categories, which include BPSK, QPSK, 8PSK, QAM16, QAM64, CPFSK, PAM4, WB-FM, AM-SSB, BFSK, and AM-DSB. It contains 11,000 signals per SNR, with each modulation category comprising 1000 samples under the length of 128. In the pre-training phase, we run 300 epochs with an Adam optimizer with a learning rate of 0.001 to optimize our model. In the fine-tuning phase, we train the downstream task heads with an Adam optimizer with a learning rate of 0.0001 about 2000 epochs, which means that the weights of the backbone in the pre-training phase are frozen at this stage. In all experiments, we utilize two distinct random seeds. For each seed, we perform ten trials and compute the average. This averaged value represents our final experimental results.

**Downstream tasks setting.** As mentioned previously, the large model for the signal can be applied to different downstream tasks. To validate the performance of the network in Figure 3, we focus primarily on the validation of the proposed model in two downstream tasks (i.e., few-shot learning (FSL) and supervised learning (SL)) in signal modulation classification. In addition, we compare the performance of different

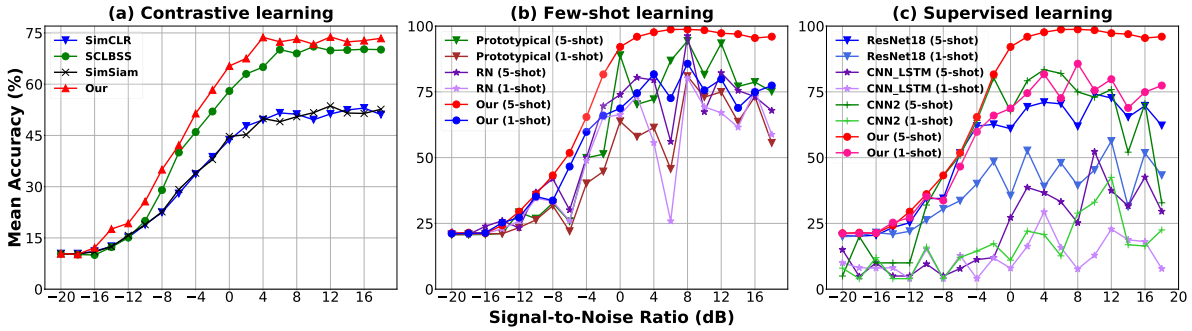


**FIGURE 4.** Experimental results on different downstream tasks on the RadioML 2016.10a dataset.

contrastive learning methods under our framework. In the FSL task, we follow the criteria in computer vision[17] to divide the signal modulation categories. Specifically, there are six categories for pre-training with 1000 samples per category. We conducted 5-way 1-shot and 5-way 5-shot classification in the fine-tuning period and all contain 500 query signals for each of the sampled categories. In the SL task, we divide the data set similarly to the FSL task. The difference is that in supervised learning, the pre-training phase has a large amount of labeled data. We compare the gap between unsupervised contrastive learning and supervised learning. It is worth noting that the sample categories for supervised pre-training and fine-tuning tests are not the same. And in contrastive learning experiments, we evaluate different contrastive learning algorithms (i.e., SimCLR,[10] SimSiam,[11] and SCLBSS[18]). For fair comparison with the previous results reported in the literature, we follow the data partitioning method of Liu et al.[18] by dividing each category in the RadioML2016.10a dataset into three parts: training, validation, and testing, with a ratio of 2:1:1. Five samples from the training set are used for fine-tuning, and the rest are used as pre-training samples. Specifically, each category includes pre-training samples, fine-tuning samples, verification samples, and test samples of 495, 5, 250, and 250, respectively.

## RESULTS

In this section, we experimentally compare several recent algorithms in contrastive learning (CL), FSL, and SL settings. Firstly, we compare three contrastive learning algorithms (SimCLR,[10] SimSiam,[11] and SCLBSS[18]) under the same data division. In Figure 4(a), at all signal-to-noise ratios, our method achieves better results compared to other algorithms. In particular, our method achieves an accuracy of 73.71% at a signal-to-noise ratio (SNR) of 4 dB, which outperforms SimCLR, SimSiam, and SCLBSS by 24.1%, 23.7%, and 8.71% in accuracy, respectively.

Then in Figure 4(b), we evaluate the pre-trained model in a downstream task of few-shot learning. Based on the above dataset partitioning criteria, the 5-way 5-shot and 5-way 1-shot tasks are tested and we have compared several classical few-shot learning networks like Prototypical Network,[19] and Relation Network (RN).[17] Our framework achieves more than 92% accuracy on a 5-way 5-shot task when SNR is above 0 dB, with a maximum accuracy of 99.09%. Compared to the other two methods, our method improves accuracy by 20% on average. It is worth noting that the contrastive methods exhibited a significant performance drop at SNR levels of -6dB and 6dB. The potential consumption could be the contaminated signal data, which disrupts the structural features of the original signal, thereby increasing the difficulty of model recognition. Importantly, while other methods merely map a few samples to the most similar feature space using prototypes and functions, they often overlook the challenge of dissimilarity. In contrast, our method specifically addresses this issue, leading to more stable model results.

Finally, we compare the proposed model with supervised learning. Our model is compared by performing supervised pre-training on a large amount of labeled data and then fine-tuned with 1-shot or 5-shot samples. In detail, we compare the CNN2, CNN-LSTM, and ResNet18 algorithms. As Figure 4(c) shows, when the SNR is higher than 0 dB, our framework has an improvement of at least 20% over the three networks compared, with a maximum improvement of 60%. It is evident that supervised learning models are only able to extract features related to known classes and are unable to generalize over new classes.

## Discussion

In essence, we tested our proposed framework on several downstream tasks. The results, as depicted in Figures 4, illustrate the strength of our method in signal extraction. This success is mainly due to our extractor which can extract features from both the time and frequency domains, offering a more comprehensive understanding compared to typical off-the-self networks. Our data augmentation strategies, such as cropping and rotation, eliminate unnecessary information and increase sample diversity, thus accelerating feature learning and improving model generalization. Although good results have been achieved in all experiments, there is still room for improvement here. The accuracy in the FSL method with a low signal-to-noise ratio needs further improvement and the performance is unstable in the 1-shot fine-tuning setting.

## CONCLUSION

The rapid development of communication technology has brought great challenges to the processing of multi-modal signals. In this paper, we made an in-depth study of large-scale models in the field of signal data and illustrated the problems, future challenges, and research directions for different stages of the development life cycle. Moreover, a preliminary study has been conducted, wherein a multitasking signal model has been proposed. This model has demonstrated decent results across multiple tasks. However, there are areas that necessitate further improvements, specifically when dealing with noisier data. Additionally, the issue of unstable accuracy during fine-tuning with a small number of samples also urgently requires a solution.

## REFERENCES


1. Z. Zhang, H. Luo, L. Zhu, G. Lu, and H. T. Shen, "Modality-invariant asymmetric networks for cross-modal hashing," *IEEE Transactions on Knowledge and Data Engineering*, 2022.
2. Z. Zhang, L. Liu, F. Shen, H. T. Shen, and L. Shao, "Binary multi-view clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1774–1782, 2018.
3. Z. Zhang, L. Liu, Y. Luo, Z. Huang, F. Shen, H. T. Shen, and G. Lu, "Inductive structure consistent hashing via flexible semantic calibration," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 4514–4528, 2020.
4. Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-Bert: Aligning image pixels with text by deep multi-modal transformers," *arXiv preprint arXiv:2004.00849*, 2020.
5. D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-E: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.
6. J. Gong, X. Xu, and Y. Lei, "Unsupervised specific emitter identification method using radio-frequency fingerprint embedded infogan," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2898–2913, 2020.
7. S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.
8. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.
9. Z. Wu, W. Cao, D. Bi, and J. Pan, "Clipc: Contrastive learning-based radar signal intra-pulse clustering," *IEEE Internet of Things Journal*, 2023.
10. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *in Proceedings of International Conference on Machine Learning*, 2020, pp. 1597–1607.
11. X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 750–15 758.
12. P. Sun, J. Su, Z. Wen, Y. Zhou, Z. Hong, S. Yu, and H. Zhou, "Boosting signal modulation few-shot learning with pre-transformation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.
13. J. Yoon, E. Yang, J. Lee, and S. J. Hwang, "Lifelong learning with dynamically expandable networks," *arXiv preprint arXiv:1708.01547*, 2017.
14. A. Ali, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek *et al.*, "XCiT: Cross-covariance image transformers." *Advances in neural information processing systems*, vol. 34, pp. 20 014–20 027, 2021.
15. K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in Neural Information Processing Systems*, vol. 29, pp. 1849–1857, 2016.
16. T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Engineering Applications of Neural Networks*,

2016, pp. 213–226.
17. 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 *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.
18. D. Liu, P. Wang, T. Wang, and T. Abdelzaher, "Self-contrastive learning based semi-supervised radio modulation classification," in *IEEE Military Communications Conference*, 2021, pp. 777–782.
19. J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4077–4087, 2017.

**Zhenyu Wen** (Member, IEEE) is currently a Professor at the Institute of Cyberspace Security and College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interests include IoT, crowd sources, AI systems, and cloud computing. Wen received his Ph.D. degree in computer science from Newcastle University. Contact him at zhenyuwen@zjut.edu.cn.

**Yuheng Ye** is currently pursuing a master's degree in computer science and technology at the School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China. His research interests include large-scale model training and fine-tuning, radio signal recognition. Contact him at 211122120010@zjut.edu.cn.

**Jie Su** is currently an assistant professor with the Institute of Cyberspace Security and College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interests include deep learning, signal processing, and the IoT security. Su received his Ph.D. degree in computer science from Newcastle University U.K., in 2023. Contact him at jieamsu@gmail.com.

**Taotao Li** is currently working toward a Ph.D. degree in control theory and control engineering with the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interests include machine learning, wireless communication security, and deep learning-based radio signal processing. Contact him at 2111903074@zjut.edu.cn.

**Jinhao Wan** is currently working toward a Ph.D. degree at Zhejiang University of Technology, Hangzhou, China. His research interests include deep learning in signal recognition and generation. Contact him at wanjinhao1@gmail.com.

**Shilian Zheng** is currently an Associate Researcher with the Science and Technology on Communication Information Security Control Laboratory, Jiaxing, China. His research interests include cognitive radio, spectrum management, and deep learning-based radio signal processing. Zheng received his Ph.D. degree in communication and information systems from Xidian University, Xi'an, China. Contact him at lianshizheng@126.com.

**Zhen Hong** (Member, IEEE) is currently a Professor with the Institute of Cyberspace Security and the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interests include the IOT, cyberspace security, and data analytics. Hong received his Ph.D. degree in control theory and control engineering from the Zhejiang University of Technology. Contact him at zhong1983@zjut.edu.cn.

**Shibo He** (Senior Member, IEEE) is currently a Professor at Zhejiang University, Hangzhou, China. His research interests include the Internet of Things, crowdsensing, and big data analysis. He received his Ph.D. degree in control science and engineering from Zhejiang University, China. Contact him at ferer@zju.edu.cn.

**Haoran Duan** (Student Member, IEEE) is currently pursuing a Ph.D. degree in the Department of Computer Science at Durham University. His research interests include applications and theories of deep learning. Duan received his M.S. degree with distinction in Data Science from Newcastle University, UK. Contact him at h.duan5@newcastle.ac.uk.

**Yuexiang Li** is currently a Senior Researcher at Tencent Jarvis Lab and has been engaged in research on intelligent analysis and processing of medical images. His research interests include microscopic images, pathological slices, and multimodal medical images. Li received his Ph.D. degree from the University of Nottingham, United Kingdom. Contact him at vicyxli@tencent.com.

**Yawen Huang** is currently a Senior Research Scientist with Tencent Jarvis Lab. Her research interests include computer vision, machine learning, and medical imaging, with a focus on practical AI for computer-aided diagnosis. Huang received her Ph.D. degree from the Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, U.K. Contact her

at yawenhuang@tencent.com.

**Yefeng Zheng** is now the Director and Distinguished Scientist of Tencent Jarvis Lab, Shenzhen, China. His research interests include medical natural language processing, computer vision, and deep learning. Zheng received a Ph.D. degree in document image analysis from the University of Maryland, College Park, USA. He is a fellow of the IEEE and AIMBE. Contact him at yefengzheng@tencent.com.