CLAP-S: Support Set Based Adaption for Downstream Fiber-optic Acoustic Recognition

Jingchen Sun[1,2 †], Shaobo Han[1], Wataru Kohno[1], Changyou Chen[2]
[1]NEC Laboratories America, Inc, USA [2]University at Buffalo, State University of New York, USA

***Abstract*—Contrastive Language-Audio Pretraining (CLAP) models have demonstrated unprecedented performance in various acoustic signal recognition tasks. Fiber-optic-based acoustic recognition is one of the most important downstream tasks and plays a significant role in environmental sensing. Adapting CLAP for fiber-optic acoustic recognition has become an active research area. As a non-conventional acoustic sensor, fiber-optic acoustic recognition presents a challenging, domain-specific, low-shot deployment environment with significant domain shifts due to unique frequency response and noise characteristics. To address these challenges, we propose a support-based adaptation method, CLAP-S, which linearly interpolates a CLAP Adapter with the Support Set, leveraging both implicit knowledge through fine-tuning and explicit knowledge retrieved from memory for cross-domain generalization. Experimental results show that our method delivers competitive performance on both laboratory-recorded fiber-optic ESC-50 datasets and a real-world dataset. Our research also provides valuable insights for other downstream acoustic recognition tasks. The code is available at https://github.com/Jingchensun/clap-s.**

***Index Terms*—Fiber-optic acoustic recognition, sound classification, domain adaptation, transfer learning, few-shot learning**

I. Introduction

Distributed acoustic sensing (DAS) [1] is a powerful technology that captures acoustic disturbances by measuring phase changes in the backscattered optical signal caused by vibrations. DAS interrogator connecting to one end of the optical fiber (spanning kilometers) transforms the cable into equally spaced sensing elements with meter-scale spatial resolution, enabling new fiber-based acoustic applications. In particular, DAS is advantageous for harsh environments such as outdoor or underwater settings, where power supply and data transmission pose challenges. Conventional sound recording device such as microphones is considered as *point sensor*, each of which only monitors one small area. In contrast, DAS can utilize long optical fibers as *linear sensors* for wide-area monitoring over vast distances without the need for numerous individual sensors and battery installation. This technology has been successfully applied to pipeline leak detection [2], rail crack detection [3], drone detection [4], utility pole localization [5], [6], seismic monitoring [7], insect activity monitoring [8], whale call detection [9], and underwater surveillance [10].

Despite high prospects in a wide range of industrial applications, developing fiber-optic acoustic recognition systems still faces some challenges: First, field data collection and annotation are labor-intensive and time-consuming. As a result, obtaining sufficient labeled data from DAS for supervised learning is often more difficult than from microphones, especially for rare classes in the long tail. Second, the characteristics of sensing data are influenced by multiple factors including sensor configuration, propagation media, signal source, and optical factors [11], [12], causing severe domain gaps. Third, users may be interested in recognizing events of new classes that are unseen during training, which leads to an open-set recognition problem.

Recently, contrastive language-audio pre-training (CLAP) models [13], [14] has emerged as a new paradigm for learning general-purpose audio representations. CLAP have demonstrated strong zero-shot performance in multiple downstream domains including sound, music, and speech. Encouraged by this success, we explore the possibility of adapting CLAP to the fiber-optic acoustic domain of interest. Due to severe domain shifts, directly using CLAP pre-trained on microphone-recorded audio data to recognize fiber-optic acoustic events results in very low zero-shot classification accuracy, e.g., less than 30% on a 50-class environmental sound dataset recorded from a fiber coil, as shown in Table II.

Existing approaches, such as Prompt Tuning [15]–[18] and Adapter methods [19]–[22], enable efficient fine-tuning on downstream tasks. Prompt Tuning leverages learnable text prompts to maximize the extraction of implicit knowledge [23], [24] from pre-trained models but often struggles to incorporate new knowledge when faced with significant domain gaps. In contrast, Adapter methods employ projection layers to explicitly acquire task-specific knowledge. Methods like Tip-Adapter [20] and Treff [25] introduce a learnable adapter to enhance CLAP models in low-shot scenarios, achieving competitive performance. However, the relative importance of implicit knowledge embedded in pre-trained models versus explicit knowledge obtained through fine-tuning remains unclear.

Is the implicit knowledge from the pre-training model always helpful? In this paper, we address the problem of adapting CLAP models for fiber-optic acoustic recognition as a downstream task. We focusing on how the pre-trained knowledge, domain shift, and limited labels will affect our adaptation process. To systematically study this problem, we create a fiber-optic version of the ESC-50 dataset [26] by replaying and recording it under various data acquisition settings in the lab. We also consider real-world gunshot vs. firework classification dataset collected from existing telecommunication cables in the field [27]. This task is particularly challenging due to differences in the recording device, the

† This work was conducted during an internship at NEC Labs.

complex outdoor environment, and the sound characteristics of the unseen and new classes.

We evaluate the effectiveness of several model adaptation approaches, including prompt tuning [15], [16], inserted adapter [19], Tip-Adapter [20] and Treff [25]. Different from existing methods, to maximize the use of labeled data, we propose fine-tuning and augmenting CLAP models with the same support set, which facilitates cross-domain generalization based on both implicit knowledge memorized in model parameters and explicit knowledge stored in the support set. The proposed approach consistently improves performance on both lab-collected dataset and field dataset.

II. Method

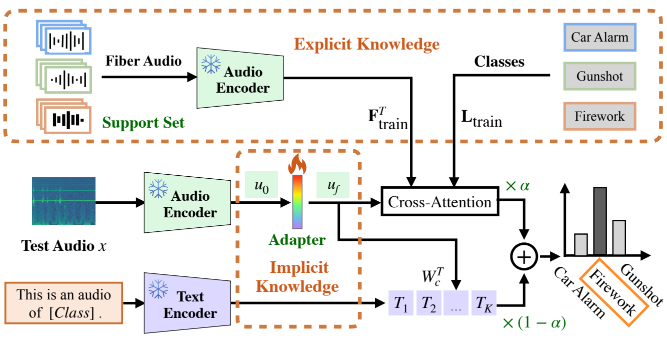

Fig. 1. **The pipeline of our proposed method.** A test sample is sent to the frozen pre-trained audio encoder and fine-tuned adapter to obtain the embedding, which is then used to perform cross-attention with the keys from the support audio samples. The attention weights are further multiplied by the values of the support set to serve as **Explicit Knowledge**. The final prediction is obtained by Linear interpolation with the Explicit Knowledge and the **Implicit Knowledge** captured by a fine-tuned Adapter .

Given a pre-trained CLAP model and a downstream dataset, we assume K-shot N-class training samples for fine-tuning. For all NK training audios X_K, the audio embeddings are represented as $F_{\text{train}} = \text{AudioEncoder}(X_K)$, where $F_{\text{train}} \in \mathbb{R}^{NK\times C}$, where C is the hidden dimension of the audio encoder. The label vectors are represented as $L_{\text{train}} = \text{OneHot}(L_N)$, where $L_{\text{train}} \in \mathbb{R}^{NK\times N}$. The audio embeddings and label vectors form the keys and values of the **Support Set**, respectively. The Support Set stores all new knowledge extracted from the training set.

For a given test audio sample x, the normalized audio embedding $u \in \mathbb{R}^{1\times C}$ is obtained by feeding the sample into the audio encoder. The audio embedding u serves as a **query**, while the stored audio embeddings F_{train} serve as **keys**. Cross-attention [28] is applied between the query and keys. The attention weights are multiplied by the **values** L^T_{train} to obtain the similarity-based prediction.

Complimentary Mechanisms of Generalization. We propose to utilize supervised information in the support set twice, for both the fine-tuning adapter and the key-value support set. Our approach uses a Linear interpolation between both implicit knowledge from model fine-tuning and explicit knowledge from the support set and seeks a balance. The final prediction is derived by

$$p_{\text{final}}(y|x) = (1-\alpha)p_{\text{clap}}(y|x,u) + \alpha p_{\text{support}}(y|x,u) \quad (1)$$

where p_{clap} is the class distribution from the fine-tuned CLAP model and p_{support} is the class distribution from the support-set-based class distribution. They are defined as follows:

$$p_{\text{clap}}(y|x,u) = uW_c^T,$$
$$p_{\text{support}}(y|x,u) = e^{-\beta\left(1-uF^T_{\text{train}}\right)} L^T_{\text{train}}.$$

Here, $\alpha \in [0,1]$ is a tuned parameter balancing the contribution from the implicit knowledge and explicit knowledge from the support set, and β is the sharpness parameter. Here u has two different representations, one is the text-aligned audio representation u_0, and the other one is the task-aligned audio representation u_{f}, they obtained by the following equation:

$$u_{\text{f}} = \text{Adapter}(u_0), \quad u_0 = \text{AudioEncoder}(x) \quad (2)$$

The Adapter is a two-layer MLP. u_0 is obtained by feeding the test audio sample x into the audio encoder, which is aligned with the text embedding during pretraining. And u_{f} is derived by passing u_0 through the two-layer MLP Adapter, while u_{f} is aligned with the task during fine-tuning.

By combining u_0 and u_{f}, we can derive many variants from Equation 1: (1) combine u_0 for p_{clap} and u_{f} for p_{support}. (2) combine u_{f} for p_{clap} and u_0 for p_{support}. (3) combine u_{f} for p_{clap} and u_{f} for p_{support}. **These combinations allow us to explore whether text-aligned or task-aligned embeddings are more effective for fiber acoustic recognition.** We empirically find that combination (3) proves to be the most effective for our task and we summarize our method two variants:

CLAP-S: Here u=u_0 and $\alpha = 1$, our final prediction relies solely on generalization through memorization via p_{support}, completely removing the influence of zero-shot or fine-tuning knowledge.

CLAP-S$^+$: Here u=u_{f} and $0 < \alpha < 1$, $p_{\text{clap-s}}$ is obtained by linearly interpolating between the knowledge in p_{clap} and the explicit knowledge p_{support} from the support set. This method uses task-aligned embeddings for both classification and retrieval. The pipeline of CLAP-S$^+$ is shown in Figure 1.

TABLE I
Implicit & Explicit Generalization Combination.

p_{clap}	p_{support}	Combination	Method	ZS
u_0		ZS (α=0)	ZS-CLAP	Yes
	u_0	Support set (α=1)	**CLAP-S**	No
u_0	u_0	ZS + Support set	Tip-Adapter	Yes
u_0	u_{f}	ZS + Support set$^{\text{F}}$	Tip-Adapter-F	Yes
u_{f}	u_{f}	Adapter + Support Set$^+$	**CLAP-S$^+$**	No

Relation with Existing works. The difference between our method and existing method is shown in Table I. Compared to Tip-Adapter [20] and Treff [25], which treats the keys in the whole support as the task-aligned embedding, our CLAP-S$^+$ take a more natural approach using embedding from the fine-tuned adapter for both query and key. This leads to stronger generalization in practice (see Table II and V).

TABLE II
ZERO-SHOT & FULL-SHOT ADAPTION ON THE FIBER-OPTIC ACOUSTIC SENSING DATASET.

	Method	Shot	Laboratory Task			Real Task		Average
			ECM	FM	FC	FMO	FCO	
Trainin-Free	[class]	zero	71.8	35.2	22.1	18.0	14.0	32.2
	this is [class]	zero	79.9	42.9	27.4	20.0	10.0	36.0
	this is an audio of [class]	zero	80.7	44.3	27.2	17.0	10.0	35.8
	i can hear the sound of [class]	zero	77.5	40.6	27.9	16.0	16.0	35.6
	Tip-Adapter [20]	full	87.0$_{\pm 0.1}$	59.0$_{\pm 0.1}$	39.0$_{\pm 0.1}$	78.8$_{\pm 1.5}$	82.2$_{\pm 1.0}$	69.2
	CLAP-S (ours)	full	92.0$_{\pm 0.1}$	61.4$_{\pm 0.8}$	43.0$_{\pm 0.0}$	79$_{\pm 0.1}$	83$_{\pm 0.1}$	71.6
	△		**+5.0 ↑**	**+2.4 ↑**	**+4 ↑**	**+0.2 ↑**	**+0.8 ↑**	**+2.4 ↑**
Training-Required	Prompt Tuning [15]	full	87.3$_{\pm 1.7}$	46.3$_{\pm 1.8}$	30.2$_{\pm 1.2}$	4.0$_{\pm 4.0}$	5.0$_{\pm 5.2}$	34.6
	Adapter [19]	full	92.9$_{\pm 1.1}$	68.8$_{\pm 1.0}$	48.8$_{\pm 1.7}$	81.4$_{\pm 0.9}$	90.2$_{\pm 0.7}$	76.4
	Treff [25]	full	90.3$_{\pm 0.8}$	67.5$_{\pm 0.9}$	46.8$_{\pm 1.0}$	84.2$_{\pm 1.3}$	90.1$_{\pm 1.7}$	75.8
	Tip-Adapter-F [20]	full	91.0$_{\pm 0.7}$	68.6$_{\pm 1.0}$	47.4$_{\pm 0.8}$	84.6$_{\pm 1.2}$	90.8$_{\pm 0.7}$	76.5
	CLAP-S^{+} (ours)	full	94.0$_{\pm 1.2}$	70.0$_{\pm 0.8}$	51.0$_{\pm 1.2}$	87.0$_{\pm 1.9}$	92.0$_{\pm 1.7}$	78.8
	△		**+2.8 ↑**	**+1.4 ↑**	**+3.0 ↑**	**+2.4 ↑**	**+1.2 ↑**	**+2.3 ↑**

TABLE III
DATASET COMPARISON.

	Dataset Name	Classes	Train	Val	Test	Total
Lab Task	Electric Microphone (ECM)	50	1400	400	200	2000
	Fiber Mandrel (FM)	50	1400	400	200	2000
	Fiber Coil (FC)	50	1400	400	200	2000
Real Task	Fiber Mandrel Outdoor (FMO)	8	647	167	421	1235
	Fiber Coil Outdoor (FCO)	8	406	106	263	775

III. EXPERIMENTAL RESULTS

We use the CLAP model from 2023 as the pre-trained model. In all subsequent experiments, each dataset is run five times, and reported the average and standard deviation in the tables. We use the AdamW optimizer with a learning rate of 1e-5, batch size of 64, and trained for 20 epochs. All experiments were conducted on an RTX A6000.

A. Datasets

Laboratory-Recorded Fiber-Optic ESC-50 Dataset To study the effects of different device domains (and recording environment), we record data using two fiber-optic sensors [29]: (1) a Fiber Mandrel (FM) with a cylinder wrapped in single-mode bare fiber [30], (2) a Fiber Coil (FC), and (3) an Electric Condenser Microphone (ECM8000) as a reference. We utilize a DAS system to record the ESC-50 dataset [26], a public benchmark dataset that includes 2000 samples from 50 categories of environment sounds.

Field Fiber-Optic Acoustic Event Classification Dataset To further evaluate the ability to distinguish fine-grained event classes and generalize to real-world environments, we consider the field fiber-optic acoustic event classification dataset [27]. It contains 8 types of real-life sound events: gunshots, crackers, cannons, fountain cannons, high-altitude fireworks, vehicle door slamming, vehicle alarms, and background noise. This dataset was collected from pre-deployed telecom networks using two types of DAS sensors in the outdoor environment: Fiber Mandrel Outdoor (FMO) and Fiber Coil Outdoor (CO). The datasets split and comparison are shown in Table III.

B. Zero-shot & Full-shot Adaptation

Zero-shot Adaptation. This scenario tests the performance of the pre-trained model when we do not have any labeled data for fine-tuning. **The results show that the domain shift on the real-world task is higher than the laboratory-recorded task.** The laboratory datasets, based on ESC-50, include typical environmental sounds (e.g., dog, engine, rain) frequently found in the pretraining data in the CLAP model under test. In contrast, the real-world dataset features uncommon categories (e.g., 'fountain cannon", a label for the firework event) and larger background noise (e.g., wind), which is considered to contribute to poor zero-shot recognition, with accuracy dropping below 20%. Besides, we also found **the domain shift in the fiber mandrel is also larger than the fiber coil** as FM achieved a higher accuracy (44.3%) than the FC (27.9%), even though they share the same label space.

Full-shot Adaptation. Once we have enough labeled data, we can inject the new knowledge from these data into the model. If the computing resource is not available, then the train-free framework is a good way. Tip-Adapter is a strong baseline in the training-free methods, significantly improving the average by 33.6 percentage points without any parameter update. However, Tip-Adapter relies on the CLAP pre-trained model, which may not help our fiber acoustic data, as a significant domain shift exists. As shown in the table, **our proposed CLAP-S, which relies solely on external knowledge retrieval from the support set, achieves better results than Tip-Adapter**. When training resources are available, fine-tuning the model can further improve the adaptation performance. **Our method, CLAP-S-$^{+}$, achieves the highest accuracy among all the baseline methods,** including Prompt Tuning, Adapter, Treff, and Tip-Adapter-F. CLAP-S-$^{+}$ utilizes task-aligned embedding to repeated use of feature knowledge and thus further improves accuracy over the Tip-Adapter-F. Prompt Tuning performed poorly on our real-world dataset, with an accuracy of only 4-5%. We hypothesize that this is due to Prompt Tuning being designed to maximize the use of pre-trained knowledge, which is less effective for tasks completely unseen by the pre-trained model.

TABLE IV
EFFICIENCY COMPARISON

Method	Acc	Training	Parameters	Time	Inference
Prompt Tuning	34.6	Required	6.14k	30min	266ms
Tip-Adapter	69.2	Free	0.0	5min	55ms
CLAP-S (ours)	71.6	Free	0.0	4min	45ms
Adapter	76.4	Required	0.52M	15min	96ms
Tip-Adapter-F	76.5	Required	1.43M	10mins	70ms
CLAP-S^{+} (ours)	78.8	Required	0.52M	7mins	56ms

C. Few-Shot Adaptation

This scenario applies when we aim to minimize labeling cost and only limited data are available. We conducted few-shot experiments ranging from 2-shot to 24-shot, as shown in Figure 2. The results demonstrate that both **CLAP-S and CLAP-S^{+} consistently improve accuracy across the four datasets and perform competitively against baseline methods.**

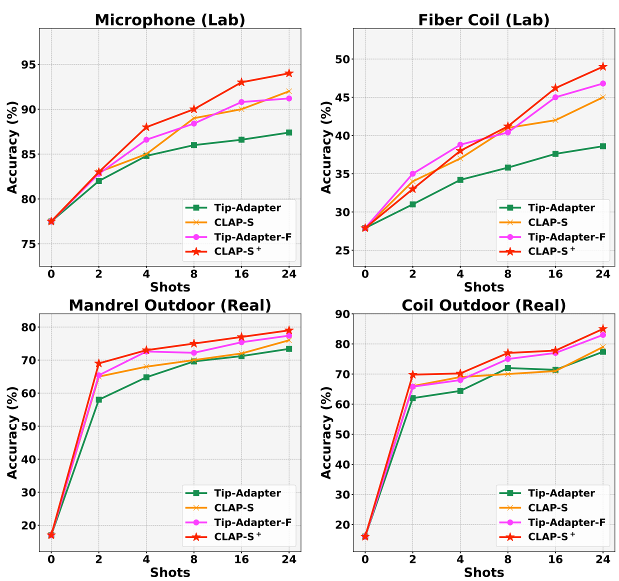

Fig. 2. **The Few-Shot Adaptation Results.**

D. Efficiency Comparison

We also compared the efficiency of our method with other baselines, as shown in Table IV. **CLAP-S demonstrated the best efficiency, requiring no training or additional parameter storage, with the lowest training and inference time.** However, this comes at the cost of sub-optimal performance. In contrast, our training-required version of CLAP-S^{+} achieves the highest average accuracy across five datasets.

E. Ablation Study

1: Does Zero-shot knowledge always contribute positively? The result in Table V shows that by adding the zero-shot ('ZS' in the table) knowledge, the performance of the Support Set model and the Adapter model both drops. Thus, **in our fiber acoustic domain, the zero-shot knowledge is not helpful**. This could be due to the misalignment of the text-audio representation caused by domain gaps and the deficiency of the language encoder in handling new acoustic concepts.

TABLE V
ABLATION STUDY

Methods	ECM	FM	FC	FMO	FCO	Avg
Support Set	$92.0_{\pm 0.1}$	$61.4_{\pm 0.8}$	$43.0_{\pm 0.0}$	$79_{\pm 0.1}$	$83_{\pm 0.1}$	71.6
ZS + Support Set	$87.0_{\pm 0.1}$	$59.0_{\pm 0.1}$	$39.0_{\pm 0.1}$	$78.8_{\pm 1.5}$	$82.2_{\pm 1.0}$	69.2
ZS + Support Set^{+}	$91.8_{\pm 1.0}$	$58.6_{\pm 1.0}$	$39.2_{\pm 1.8}$	$25.0_{\pm 0.1}$	$20.0_{\pm 0.1}$	46.9
Adapter	$92.9_{\pm 1.1}$	$68.8_{\pm 1.0}$	$48.8_{\pm 1.7}$	$81.4_{\pm 0.9}$	$90.2_{\pm 0.7}$	76.4
Adapter + ZS	$94_{\pm 1.1}$	$68.6_{\pm 0.5}$	$48_{\pm 1.1}$	$82.2_{\pm 1.3}$	$85.8_{\pm 2.1}$	75.7
Adapter + Support Set	$93.4_{\pm 1.4}$	$71.6_{\pm 1.0}$	$50.6_{\pm 1.0}$	$84.0_{\pm 1.5}$	$88.3_{\pm 1.3}$	77.6
Adapter + Support Set^{+}	$\mathbf{94.0}_{\pm 1.2}$	$\mathbf{70.0}_{\pm 0.8}$	$\mathbf{51.0}_{\pm 1.2}$	$\mathbf{87.0}_{\pm 1.9}$	$\mathbf{92.0}_{\pm 1.7}$	**78.8**

This conclusion applies only to the specific domain considered, yet one can see that when significant domain gaps exist and acoustic events are difficult to describe with language, the zero-shot transfer may have unintended negative effects.

2: Which representation is more effective for retrieval in the support set — text-aligned or task-aligned? The text-aligned embedding for the Support Set is represented as 'Support Set', while the task-aligned embedding for the Support Set is represented as 'Support Set^{+}' in Table V. The results indicate that, for the same ZS or Adapter, adding 'Support Set^{+}' consistently yields better performance than adding the 'Support Set'. **Thus the task-aligned representation is more effective for key-value retrieval** than text-aligned representation. This is due to the larger domain gaps encountered in the fiber-optic acoustic domain than the conventional microphone domain [25].

3: Fine-tuning jointly or separately? In scenarios with data collected from multiple devices or neighboring channels, should we fine-tune one model or individual models separately? We conducted experiments and the results show **the model trained jointly on all the datasets outperformed the models trained independently** on each individual dataset (results shown in Table VI). The reason is that joint training serves as a form of data augmentation, where increased data improves performance on individual tasks.

TABLE VI
JOINTLY TRAINING VS INDEPENDENTLY TRAINING

Methods	ECM	FM	FC	FMO	FCO	Avg
Adapter-Independently	$92.9_{\pm 1.1}$	$68.8_{\pm 1.0}$	$48.8_{\pm 1.7}$	$81.4_{\pm 0.9}$	$90.2_{\pm 0.7}$	76.4
Adapter-Jointly	$\mathbf{93.4}_{\pm 0.9}$	$\mathbf{70.8}_{\pm 1.5}$	$\mathbf{50.2}_{\pm 1.6}$	$\mathbf{81.9}_{\pm 0.7}$	$\mathbf{90.9}_{\pm 0.5}$	**77.4**

IV. CONCLUSION

In this paper, we explored efficient model adaptation methods by balancing implicit knowledge and explicit knowledge for the challenging fiber-optic acoustic recognition task. Our work provides valuable insights into utilizing pre-trained models to design domain-shift mitigation strategies, and improve the robustness of text-aligned and task-aligned representations. Our method may also work for other scenarios with significant domain shifts from the pre-trained model, not limited to the fiber-optic domain. Extending our approach to other downstream tasks would be a interest research direction, which we leave for future work.

References

[1] Ezra Ip, Fabien Ravet, Hugo Martins, Ming-Fang Huang, Tatsuya Okamoto, Shaobo Han, Chaitnaya Narisetty, Jian Fang, Yue-Kai Huang, Milad Salemi, et al., "Using global existing fiber networks for environmental sensing," *Proceedings of the IEEE*, vol. 110, no. 11, pp. 1853–1888, 2022. 1

[2] Femi Tanimola and David Hill, "Distributed fibre optic sensors for pipeline protection," *Journal of Natural Gas Science and Engineering*, vol. 1, no. 4-5, pp. 134–143, 2009. 1

[3] Cunzheng Fan, Fan Ai, Yijie Liu, Zhijie Xu, Geng Wu, Wei Zhang, Chen Liu, Zhijun Yan, Deming Liu, and Qizhen Sun, "Rail crack detection by analyzing the acoustic transmission process based on fiber distributed acoustic sensor," in *2019 Optical Fiber Communications Conference and Exhibition (OFC)*. IEEE, 2019, pp. 1–3. 1

[4] Jian Fang, Yaowen Li, Philip N Ji, and Ting Wang, "Drone detection and localization using enhanced fiber-optic acoustic sensor and distributed acoustic sensing technology," *Journal of Lightwave Technology*, vol. 41, no. 3, pp. 822–831, 2022. 1

[5] You Lu, Yue Tian, Shaobo Han, Eric Cosatto, Sarper Ozharar, and Yangmin Ding, "Automatic fine-grained localization of utility pole landmarks on distributed acoustic sensing traces based on bilinear resnets," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4675–4679. 1

[6] Zhuocheng Jiang, Yue Tian, Yangmin Ding, Sarper Ozharar, and Ting Wang, "Utility pole localization by learning from ambient traces on distributed acoustic sensing," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5. 1

[7] María R Fernández-Ruiz, Marcelo A Soto, Ethan F Williams, Sonia Martin-Lopez, Zhongwen Zhan, Miguel Gonzalez-Herraez, and Hugo F Martins, "Distributed acoustic sensing for seismic activity monitoring," *Apl Photonics*, vol. 5, no. 3, 2020. 1

[8] Sarper Ozharar, Jessica L Ware, Yue Tian, and Yangmin Ding, "Long-term monitoring and analysis of brood x cicada activity by distributed fiber optic sensing technology," *Journal of Insect Science*, vol. 23, no. 6, pp. 3, 2023. 1

[9] Léa Bouffaut, Kittinat Taweesintananon, Hannah J Kriesell, Robin A Rørstadbotnen, John R Potter, Martin Landrø, Ståle E Johansen, Jan K Brenne, Aksel Haukanes, Olaf Schjelderup, et al., "Eavesdropping at the speed of light: Distributed acoustic sensing of baleen whales in the arctic," *Frontiers in Marine Science*, vol. 9, pp. 901348, 2022. 1

[10] Bin Lu, Bingyan Wu, Jinfeng Gu, Junqi Yang, Kan Gao, Zhaoyong Wang, Lei Ye, Qing Ye, Ronghui Qu, Xiaobao Chen, et al., "Distributed optical fiber hydrophone based on ϕ-otdr and its field test," *Optics express*, vol. 29, no. 3, pp. 3147–3162, 2021. 1

[11] Jun Zhou, Zhengqing Pan, Qing Ye, Haiwen Cai, Ronghui Qu, and Zujie Fang, "Characteristics and explanations of interference fading of a ϕ-otdr with a multi-frequency source," *Journal of Lightwave Technology*, vol. 31, no. 17, pp. 2947–2954, 2013. 1

[12] Noriyuki Tonami, Wataru Kohno, Sakiko Mishima, Yumi Arai, Reishi Kondo, and Tomoyuki Hino, "Low-rank constrained multichannel signal denoising considering channel-dependent sensitivity inspired by self-supervised learning for optical fiber sensing," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8511–8515. 1

[13] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5. 1

[14] Benjamin Elizalde, Soham Deshmukh, and Huaming Wang, "Natural language supervision for general-purpose audio representations," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 336–340. 1

[15] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022. 1, 2, 3

[16] Yiming Li, Xiangdong Wang, and Hong Liu, "Audio-free prompt tuning for language-audio models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 491–495. 1, 2

[17] Asif Hanif, Maha Tufail Agro, Mohammad Areeb Qazi, and Hanan Aldarmaki, "Palm: Few-shot prompt learning for audio language models," *arXiv preprint arXiv:2409.19806*, 2024. 1

[18] Jingchen Sun, Jiayu Qin, Zihao Lin, and Changyou Chen, "Prompt tuning based adapter for vision-language model adaption," *arXiv preprint arXiv:2303.15234*, 2023. 1

[19] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024. 1, 2, 3

[20] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li, "Tip-adapter: Training-free clip-adapter for better vision-language modeling," *arXiv preprint arXiv:2111.03930*, 2021. 1, 2, 3

[21] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5. 1

[22] Samuel Kessler, Bethan Thomas, and Salah Karout, "An adapter based pre-training for efficient and scalable self-supervised speech representation learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3179–3183. 1

[23] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis, "Generalization through memorization: Nearest neighbor language models," *arXiv preprint arXiv:1911.00172*, 2019. 1

[24] Zitong Yang, Michal Lukasik, Vaishnavh Nagarajan, Zonglin Li, Ankit Rawat, Manzil Zaheer, Aditya K Menon, and Sanjiv Kumar, "Resmem: Learn what you can and memorize the rest," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 1

[25] Jinhua Liang, Xubo Liu, Haohe Liu, Huy Phan, Emmanouil Benetos, Mark D Plumbley, and Wenwu Wang, "Adapting language-audio models as few-shot audio learners," *arXiv preprint arXiv:2305.17719*, 2023. 1, 2, 3, 4

[26] Karol J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. pp. 1015–1018, ACM Press. 1, 3

[27] Shaobo Han, Ming-Fang Huang, Tingfeng Li, Jian Fang, Zhuocheng Jiang, and Ting Wang, "Deep learning-based intrusion detection and impulsive event classification for distributed acoustic sensing across telecom networks," *Journal of Lightwave Technology*, 2024. 1, 3

[28] A Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017. 2

[29] Wataru Kohno, Noriyuki Tonami, Jian Fang, Shaobo Han, Jingchen Sun, and Ting Wang, "Text-guided device-realistic sound generation for fiber-based sound event classification," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025. 3

[30] Wataru Kohno, Jian Fang, Shuji Murakami, Giovanni Milione, and Ting Wang, "Distributed fiber-optic sensor as an acoustic communication receiver array," *28th International Conference on Optical Fiber Sensors*, p. Tu3.33, 2023. 3